Retailing Analysis Using Hadoop and Apache Hive

Hiba A. Abu-Alsaad

Computer Engineering Department
Faculty of Engineering, Mustansiriyah University
Iraq, Baghdad.

E-mail: Eng.hibaakram@uomustansiriyah.edu.iq

***Abstract* - Convenience is an important factor for people in daily activities particularly for users who consume goods and services and also for retailers who provide such services. The retail industry took infant steps in the early 20th century across most of Europe and America. However, there was a considerable surge in the rise of supermarkets and hypermarkets in 2nd half of the century as they provided a convenient all-in-one-stop experience for customers. This subsequently created a huge growth of data in retail stores which presented a challenge for store owners to interpret with traditional business intelligence tools. Therefore, there was a need for real-time analytic tools to handle large datasets in sizes of up to Terabyte magnitudes. Query languages such as Hive and Pig became prominent in the analysis of customer data to ensure continued convenient experience for customers and quality provision of services for retailers. In this paper we analyze the underlying factors that have made Hive to be effective for the retail industry.**

Keywords - Hive; Retail industry; Big data; Query; Database; Real-time analysis; Market.

I. INTRODUCTION

Today, retailing is a pretty massive industry and will definitely remain that way in the anticipated future. Sales, world-wide, are being estimated to be at about 7.5 trillion US Dollars [1]. According to Peter Carbonara [2], Walmart is still the biggest retailer in the world and ranked number 1 amongst other retailers on the list of *Forbes Global 2000*. Other top retailers include *Amazon*, *CVS*, *Alibaba*, *Carrefour* (in European countries like France and Romania), *Royal Ahold* (Dutch Retail Company), etc. It's also becoming big in developing countries like *Shoprite* in Nigeria, and *Tata Group* and *Aditya Birla Group* in India.

Despite what you say about retailing giants (such as Walmart, Amazon, Carrefour, etc.), there will always be a place for local stores or regional chains, which are well-run with excellent service. With more possibilities for online marketing, it creates more avenues for retail shops to thrive. However, retailing in this present century presents an exciting, yet complex, sector of the economy of most developed and developing nations. The business of retailing is being barraged by a number of forces which include an explosion in the availability of customer data, the emergence of Radio Frequency Identification (RFID) technology, growing competition amongst and across diverse retailing formats, online retailing, etc. [3] To make sense of all the forces, whilst still providing quality and convenient services to customers, is not an easy task, but one which is of extreme importance to analysts, retailing gurus, and those who make policies.

One of the big trends with which these stakeholders have to grapple with is the creation of voluminous customer data by ubiquitous devices locally and online. Such data can be rightly described as big data, which is characterized by the V's of variety (from diverse sources), velocity (the fast pace at which it accumulates), veracity (the uncertainty of data), and volume (the rate at which it scales exponentially).

Prior to the advent of big data, customers could only get personalized experience through loyalty programs [4], however, big data has become a big game changer to the possibilities it comes with the retailing industry. When it comes to getting optimum benefit from big data, it's not what how much do you have that counts; what do you do with but what you have is of utmost paramount.

Novel sources of data ranging from log files and trade information, to sensor data and social media metrics, present a wide range of opportunities for retailers to actualize unprecedented value and competitive advantage in a sector of the economy which is ever growing.

In this paper we propose methods of optimizing customer experience and maximizing profits by analyzing data generated from interactions between customer and stores by using Hive built on *Apache*. In Section 2, we briefly mention some studies related to developments aimed at improving retailing by using big data analysis. Section 3 describes the changing relationship between data analysis and retailing and describes some use-cases such as how big data analysis can be in retailing to a big game changer. In Section 4, the implementation of our work and the summary of the results obtained in a one-week period are described. Finally, Section 5 describes our conclusions and recommendations.

II. RELATED WORK

Analyzing historical data is a trending part of business management nowadays. There have been many research studies dedicated to such tasks.

For instance, the following studies were conducted on Big data analysis using Hadoop and Hive in different areas. Mehta J. and Woo J., used Hive that conducted a financial Analysis for New York Stock Exchange (NYSE) historical data for over 14 years to identify top companies using [5]. Belarbi H. and Bennis H. [6] describes various tools that can be used for big data analysis in the retailing industry such as data warehouse, distributed systems, Extraction, Transformation, and Loading (ETL) systems, and Hadoop. In another paper [7], Bradlow E. T., et. al looks at several other avenues for exploiting big data analysis in retailing. They considered analysis data arising from products, time, customers, location (geographic data and information) and channels. In Essentials of Business Analytics, the writers describe three major categories of business analysis [8], which include descriptive analysis, prescriptive, and predictive analysis. They argue that these provide the most advanced methods for providing the best course of actions for businesses to take. In a study [9] by Emel Aktas E. and Meng Y. investigated big data applications are employed in retail logistics in areas such as availability, pricing, assortment, and layout planning; they also proposed a profile for retail businesses to optimize their retail online experiences.

III. RETAILING AND BIG DATA

For a stakeholder in retailing to triumph which is a game of little successes, because, most times, retail margins are so small, and ensuring profitability, whilst taking delivery is costs and expensive and moreover is imperative. Inferring useful information from big data has the possibility of answering questions such as how to understand customer sentiments and what their desires are. To answer these questions would ensure new customers are attracted while keeping (and building) the loyalty of old customers. By analyzing big data streams (from sales, income, operations, stock, and other sources) efficiently (and in possible real-time), retailers can restructure their operations to ensure costs are reduced, consumer satisfaction is boosted, and more income is yielded [10].

According to reports by McKinsey, companies that invest in big data analysis, over a 5-year period, as part of their marketing and sales programs yield a Return on Investment (ROI) of 15-20%. Stakeholders in retail industry, in recent times, using big data has been to improve operations such as marketing, store management, vending, supply chain, online commerce and multichannel [10].

We shall consider some using cases where big data can change the game in the retailing industry.

A. Providing Personalized Experience for Customers in Stores

Novel means to analyze the behavior of customers in stores and measure the influence of marketing efforts is being guaranteed by the rise of human-tracking technologies. A platform for data analysis can aid retailers get useful information from the datasets they collect in order to optimize tactics for sales, provide right and timely promotions as incentives for customers to make complete purchases, and personalize the experience consumers have in stores with apps that boost customer loyalty. The final aim of these activities would be to raise sales across difference channels from online stores to physical stores from the insights derived from data engineering [11].

These insights can be collected from mobile apps, Point of Sale (POS) systems, store sensors, in-store cameras, websites, supply chain systems, etc.

With these platforms and insights, retailers who are both online and having physical stores can do the following:

1- Test the effect of marketing schemes on consumer behavior.

2- Closely observe customer behavior and offer timely offers as incentives to customers to make later purchases in stores or online, retaining customer loyalty in the course.

3- Utilize a customer's online trading activities and history to predict the needs and interests, thereby creating the opportunity to personalize shopping experiences for customers when they come to physical stores.

B. Improved Operations and Decisions

Data engineering can enable retailers to better comprehend chains of supply and product distribution to lower overhead costs. Utilizing big data analytics to improve efficiency of operations is being able to use them to unearth insights which are hidden in sensor, machine, and log data, which might be structured, semi-structured, or unstructured.

Assets that could churn out valuable data include company servers, appliances owned by customers, cell towers, energy grid infrastructure, transaction logs, plant machinery, etc. Insights from these data streams can provide information such as trends and patterns that can create better performance for operations, improve decision-making, and help retailers save huge amounts of money [11].

C. Making the best of Customer Behavior Analysis

The challenge retailing in this century poses is that customers, today, have the possibility of transacting and interacting via multiple points of service, which include social media platforms, physical stores, e-commerce sites, smart devices, etc. This creates a complex and varied stream of data sources to gather together and analyze.

However, with proper data engineering platforms, important insights such as these can be obtained: who are a

retailer's high-value consumers, what incentive makes them to make more transactions, what are their behaviors and idiosyncrasies, and what time is best to reach out to them [11].

D. Using Targeted Promotions to Boost Conversion Rates

Promotions are one of the surefire ways to increase conversion rates. However, to ensure high customer acquisition and lower expenses for these promotions, retailers need to ensure customer promotions are targeted effectively.

The availability of ubiquitous internet-enabled smart devices and multiple social platforms creates a culture of customers, probably, interacting more than they are having transactions. Previously, customer information would have to be gotten from demographic data gotten during customer and retailer interactions and transactions. However, interactions in this century occur massively on social media platforms through diverse channels. Therefore, it will be of paramount importance for retailers to glean customer information and insight from the massive amount of data generated by these online interactions.

With these platforms and insights, retailers who are both online and having physical stores can do the following:

1- Observing social media activity of customers and purchasing behavior to make offers to customers and to make later online purchases or requests for them in-store purchases.

2- Examining the effect of diverse promotional schemes on the behavior of customers and quantify the rate of conversion in transaction value.

3- Personalizing promotions for consumers by identifying specific needs and interests from consumer online purchasing activity.

IV. IMPLEMENTATION AND RESULTS

The implementation of this project was accomplished through a series of steps. The first step involved building an online grocery web application that offers ordering services for different products like most e-commerce applications. The second step was to collect user data and run analysis on it using Hive. The goal is to extract useful ideas and trends about customer activity during transactions, which help to improve the business. We will limit the detailed description to data analysis section, as building a web application is outside the scope of this paper.

A. Main E-commerce website

Research that concerns with data analysis comes with results that have benefits of certain area; ours are not an exception. We built an e-commerce application that provides an online shopping experience for people's daily food needs. The application was built with *Java EE* platform [13] and a snippet of the application interface is shown in Figure (1) below:

Figure 1. E-commerce web Application.

B. Dataset

The data we used for analysis was acquired from the application's Relational Database Management System (RDBMS), as we have access to the application's repository. For purposes of this research, our test data was limited to a period of one week for one location. The main reason for the test-data size was due to limited hardware resources for processing huge datasets.

C. Environment

For analysis, there were many preliminary settings that needed to be complete before proceeding to actual code development. The first step was to install *Hadoop* and install *Apache Hive* on top of it. For our research purposes, we used *Cloudera* virtual machine [14]. This workaround saved us a lot of time, as we could skip all required installations and configurations if we had to install from scratch. This was appropriate for code-testing a prototype in a small environment. However, this method will not be effective in large industrial environment as scalability might pose an issue for the processing equipment's.

1) Hive: Analyzing large datasets won't be possible without taking into consideration the efficiency of the technology in use. As the data continues to grow, we should be able to process it in an efficient and scalable manner. This can be done with the use of modern tools to avoid sophisticated and expensive equipment, which is necessary sometimes. Apache Hive is a concrete example of a data-warehouse application that provides ease of use to read, write, and manage enormous datasets located in distributed storage by using SQL language. Implement queries is established by the command line interface, using Java Database Connectivity (JDBC) connector to connect developers to hive.

Hive is used to extract some results about the products, sales, and profit.

The first search was to find the top three best-selling products. Then, results are generated for the top three products that generate the highest income to the business. Finally, the top three zones that rake in the best sales are generated.

The database in the web application has been imported to the warehouse from a MySQL database. Next, the Hive script runs on the tables and makes new relationships to calculate extract results. The data stored in HDFS and all the processing operations is on that version.

D. Workflow

Figure (2), below, gives a descriptive chart that demonstrates the interactions between the elements in order to achieve the final results.

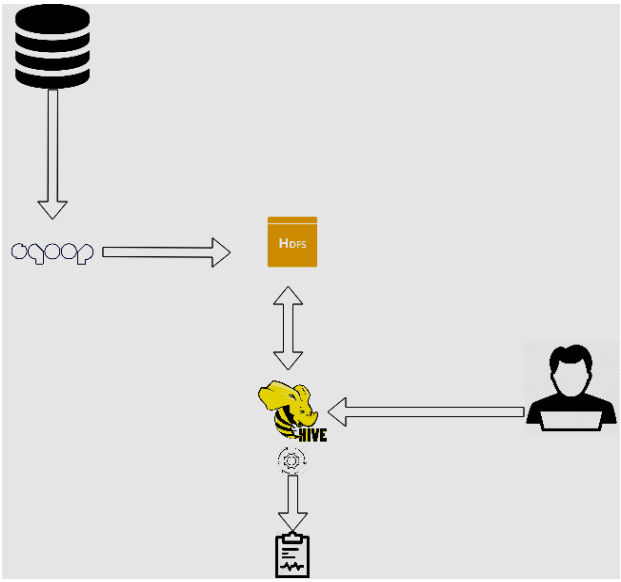

Figure2. Workflow Chart.

From the diagram, customer interactions are collected and stored in a relational database. For the possibility of parallel processing, sqoop is responsible for moving the dataset from database to HDFS. When the data is stored on HDFS, the data analyst can access it to make queries and analysis to generate results that will be used for improving business logistics and operations.

E. Data Summary

As we mentioned earlier on, the hardware equipment for analysis were limited, hence, the results are obtained in quite a reasonable time. Our test is performed on a single machine and utilized only one chunk of data. The tables below summarize results from our hive queries.

Table I represents the top three results of products which sell best in a descending order. Table II represents products which generated the most income in descending order. Finally, Table III summarizes the top 3 zones which raked in the most income during the time period.

TABLE I. BEST-SELLING PRODUCTS.

Sold-Products	Highest Number of Sales
Butter	41
Cheese	30
Free range eggs	30

TABLE II. BEST PROFITABLE PRODUCTS.

Sold-Products	Highest Income – Generation per Day
Parma ham	94.23 $
Sausages	78.1 $
Cheese	71.7 $

TABLE III. BEST-SELLING ZONES.

Zones	Highest Sales
Baghdad, Bayaa	124.22
Baghdad, Karkh	113.13
Baghdad, Al Rashid	100.34

V. CONCLUSIONS

The advantages of modern-day data analysis for e-commerce applications provide outstanding opportunities to both new and existing retailing businesses.

There are two important factors that make big data analysis the most significant leap in data analysis history. The availability and low cost of hardware commodity make big data analysis the current trend of this era. Huge datasets amount to terabytes and require very complex operations and algorithms to run on it, yet the timeframe of achieving results are in minutes, sometimes even seconds. The equipment and tools to accomplish such results are affordable and easily-obtainable to most industrial corporations, and it does not require any special hardware.

This paper proposed a way to analyze and improve retailing businesses by using big data tools, particularly *hive*, for stacked data analysis. That would give motivation to most entrepreneurs to open an online interface in addition to their in-store interactions with customers and all the necessary information will be easy to obtain by using this technique. Compared to other ways of large data analysis,

using analysis with Hadoop takes minimum time, minimum technical experience, and the most feasible way for retail stores to analyze data.

Currently any e-commerce website demands powerful advertising campaigns in order to be successful. Without publicity, there is almost no chance for considerable success for web application or websites to effectively accomplish its purpose, because of content-saturation on the web.

According to *Take Eye* website, in their report in January 2018, there are 1,805,260,010 websites in the World Wide Web [12], the potential of making well- known e-commerce website is relatively low without the proper marketing and customer targeting. For that reason, analysis related to advertising is our main next goal at this research, such as *Google Analytics* and *Facebook Ads*. Analyzing such data can give us clear insight into the people reached through the website and find weak points in the products or the advertising strategies.

Another aspect that can be improved is to build a real-time analysis by using other technical tools in order to gain more efficiency in updating the results.

ACKNOWLEDGMENT

The author would like to thank Mustansiriyah University (www.uomustansiriyah.edu.iq) Baghdad – Iraq for it is support in the present work.

REFERENCES

[1] Ganesan H., Vijay. Hive for Retail Analysis, April 24, 2012. Visited August 2018 from YourStory: https://yourstory.com/2012/04/hive-for-retail-analysis/

[2] Carbonara P. Walmart, Amazon Top World's Largest Retail Companies, June 6, 2018. Visited August 2018 from Forbes: https://www.forbes.com/sites/petercarbonara/2018/06/06/worlds-largest-retail-companies-2018/#2c454b8713e6

[3] Krafft M., Mantrala M. Retailing in the 21st Century: Current and Future Trends, January 2010. Visited August 2018 from ResearchGate: https://www.researchgate.net/publication/321614572_Retailing_in_the_21st_Century_Current_and_Future_Trends

[4] - Patel N. How 4 Major Retailers are Using Big Data. Visited August 2018 from NeilPatel: https://neilpatel.com/blog/retailers-are-using-big-data/

[5] Jay Mehta, Jongwook Woo." Big Data Analysis of Historical Stock Data Using HIVE", ARPN Journal of Systems and Software, ISSN 2222-9833, August ,2015.

[6] Hamza BELARBI, Abdelali TAJMOUATI, Hamid BENNIS, Mohammed EL HAJ TIRARI." Predictive Analysis of Big Data in Retail Industry: Literature Review", 1st International Conference on Computing Wireless and Communication Systems (ICCWCS-2016), ISSN: 2509-2014, November ,2016.

[7] Eric T. Bradlow , Manish Gangwar, Praveen Kopalle , Sudhir Voleti. "The Role of Big Data and Predictive Analytics in Retailing", Journal of Retailing 93(1):79-95, DOI: 10.1016/j.jretai.2016.12.004, March, 2017 .

[8] Camm J, Cochran J, Fry M, Ohlmann J, Anderson D."A Categorization of Analytical Methods and Models. In Essentials of Business Analytics ", pp. 5 – 7, August, 2014.

[9] Emel Aktas, Yuwei Meng." An Exploration of Big Data Practices in Retail Sector", doi:10.3390/logistics1020012, December, 2017.

[10] 6 Big Data Use Cases in Retail. Visited September 2018 from Ingram Micro Advisor: http://www.ingrammicroadvisor.com/data-center/6-big-data-use-cases-in-retail

[11] Hitchcock E. Five Big Data Use Cases For Retail, February 27, 2018. Visited July 2018 from Datameer: https://www.datameer.com/blog/five-big-data-use-cases-retail/

[12] Fowler D. How Many Websites Are There In The World?, July 1 2014, Visited July 2018: https://tekeye.uk/computing/how-many-websites-are-there

[13] Oracle Comparision between Java SE and Java EE, Visited August 2018: https://docs.oracle.com/javaee/6/firstcup/doc/gkhoy.html

[14] Cloudera QuickStarts for CDH 5.13, Visited August 2018: https://www.cloudera.com/downloads/quickstart_vms/5-13.html