Exposing Retail Trends with the Hadoop, Spark, and Apriori

Name: Saketh Reddy Atla

Mail Id: x22218700@student.ncirl.ie

Student Id: 22218700

Department Name: Data Analytics

Name of the Subject: Scalable Systems Programming

ABSTRACT

This research is going to explore the field of retail dataset, the subject we are going to navigate using Hadoop, Spark and Apriori algorithm. Aimed at enhancing comprehension of customers by the businesses and helping them make clever choices is our target. Through Hadoop's powerful data processing capabilities and Spark's multitasking operations, we can dig through large transactional databases effortlessly. In this case, we can highlight the hidden information that are deep inside the data set, such as the common products that are bought from the same consumer group. Through Apriori algorithm we can detect common patterns as well as conditions defined in the data. We will have this outcome when our consumers who shop for milk may also do that with bread. To facilitate our information presentation, we use Matplotlib script to produce easily readable charts with relevant key points. This consists of statistics and plots that carry forth story of the data with these features. Through the integration of these technologies' businesses are equipped with the instruments required to achieve a full picture of their clients, build marketing strategy aiming at consumers' interests, and consequently strengthen them.

MapReduce, Spark, Hadoop, Patterns, Apriori algorithm, Multitasking, Matplotlib

1. INTRODUCTION

The world of shopping has become a space where you can uncover the mysteries about the items we purchase. Whether it is the reason why people buy specific items or the methods to cope with the evolving patterns in the business sphere, this is all information that anybody can extract from the data produced by retail purchasing by consumers.

We examine the customer motivational forces that determine the choices they make. With stores battling for clients' mention and customers increasingly anticipating more than they were getting before, it is important for businesses to know just what makes customers choose one shop over the other.

We have got two criteria to base our analysis. Firstly, by finding patterns and links that others have failed to see in retail data, we want to outperform everyone else. Through the process of discovering what products sell best when bundled as a set or through certain customer's behaviors, we can offer businesses higher quality decisions about what to sell and how to address the selling process. In our second point, we are planning to transition the results of these insights into specific strategies that businesses can implement to enhance their operations and

delighted their customers more by doing so.

To guide our exploration, we can reduce these issues to a couple of questions that will help us to realize how shopping is done nowadays and how businesses can respond. We are always hunger, everything from which product s are sold more amidst certain seasons, to how different types of customers want to shop.

On the other hand, in this paper we further explain how we find the answers for the given questions. We will tell you about our data sources, how we made sense of it, and how we took care of some ethical issues. As such, we will demonstrate the different types of tools and methods that we applied to make out of that data. Through these findings we want to enable academics, businesses, and others to get a more genuine comprehension of the sphere of retail.

Research Question

Which items are the mostly ordered together?

Are we able to differentiate clients according to their purchasing patterns?

Will we be able to offer additional products to the customers by reference to their actual purchase behaviour?

Research Objectives

The research objectives attained by the implementation of Market Basket Analysis, Customer Segmentation, and Product Recommendation in Spark will vary according to the details and targets of the research job. undefined

Understanding Customer Behavior: Put through transactional data to identify trends in customer purchasing patterns, e.g. frequently bought together items, segmentation of customers based on their purchasing preference and preference tendency based on demographics, gender, etc.

Optimizing Marketing Strategies: Apply information obtained from market basket analysis and customer segmentation to refine marketing strategies by sending promo letters, personalized suggestions, and multi-product bundle offers, generate an increment in profit and happy customers.

Enhancing Product Recommendations: Create and improve recommendation systems that suggest personalized products/services to clients based on their past purchase behavior, by doing so customer engagement and retention grows.

Identifying Cross-Selling Opportunities: Finding occasions for cross selling and upselling through analyzing of the association rules by presenting the customers the relevant products by their shopping journey.

Improving Operational Efficiency: Streamline inventory management and supply chain systems by locating products which are most liked, determine demand, and organize stocks based on historical sales data and wishes of customers.

2. Literature Review

In 2008, Li [1] proposed an association rule mining algorithm called as PFP (Parallel Frequent Pattern). This algorithm is a parallel implementation of FP-Growth (Frequent Pattern-Growth) algorithm based on MapReduce paradigm. It eliminates the requirements of data distribution and load balancing by using

MapReduce paradigm. It was highly scalable and quite suitable for web data mining. PFP search for top-k patterns instead of patterns fulfilling user specified minimum support criteria which make it effective for web data mining. Author applies it on query log for search recommendations. PFP load balancing technique was not so efficient, therefore Zhou [2] proposed a new algorithm called as BPFP (Balance Parallel FP-Growth). BPFP algorithm has better load balancing technique to make PFP faster and efficient.

In 2012, Li [3] proposed a MapReduce-based association rule mining algorithm and ran it on Amazon EC2 cluster. This algorithm uses basic apriori approach. It provides faster results due to high computation power available on EC2. It uses Amazon S3 for data storage. Later, lin [4] proposed three MapReduce-based association rule mining algorithms called as SPC (Single Pass Counting), FPC (Fixed Passes Combined-counting), and DPC (Dynamic Pass Counting). SPC algorithm is simple MapReduce implementation of Apriori. FPC algorithm works same as SPC for finding up to 2-itemsets. It combines candidate sets for remaining passes to get results in a single phase. It was useful in some cases where the size of candidate set is small after two iterations and many machines in Hadoop cluster remain idle during afterward steps. By default, FPC combines candidate set of 3 iterations like candidate set of 3-itemsets, 4-itemset and 5-itemset. FPC also has a drawback that it combines fix number of passes and have the possibility of crashing if candidate set for higher iterations is large. DPC algorithm resolve this issue quite efficiently by combining phases depending on candidate size and machines computation power. It provides better load balancing for increasing efficiency. In the same year, Li [5] and Yahya (MRApriori) [6] proposed another MapReduce paradigm-based association rule mining algorithms which are a straightforward parallel implementation of Apriori algorithm

In 2014, Lin [7] came with one more MapReduce paradigm based parallel implementation of Apriori. Lin uses most outdated cluster hardware (P4) for computations. Later, Barkhordari [8] proposed a MapReduce-based association rule mining algorithm called as Sca DiBino (scalable and distributable binominal association rule mining algorithm) which converts every row of input transactions to a binomial format. Binomial data can be processed more efficiently on MapReduce. This algorithm directly generates association rule without finding frequent patterns. They used this algorithm for recommending value added services to customers by analyzing network traffic of a mobile operator. Some researchers use various data structures to improve the efficiency of association rule mining algorithms. Singh [9] tries to use a hash table, hash tried, and hash table tried for candidate storage in Apriori MapReduce-based implementation. They find that hash table tried is most efficient than others in MapReduce context while it is not much efficient in a sequential approach.

3. Methodology

Description of Data Sets and Attributes:

Dataset has been downloaded from OpenML which consists of many datasets for applying machine learning algorithms. The data set ranges from the transaction records of a retail business which contain characteristics such Invoice No, Stock Code, Description, Quantity, Invoice Date, Unit Price, CustomerID, and Country.

Translation Rules for Data Manipulation:
The process of developing the necessary rules for

translation is carried out for facilitating data handling. Included in such strategies are procedures for filling the empty and null values, attribute splitting or joining, and the mapping process for a standardization.

Architecture and Application Workflow Overview:

A distributed processing capability is provided by the joint implementation Hadoop MapReduce and Apache Spark computing framework oriented for large data processing. In the first place, the Hadoop MapReduce works on preprocessing and shaping data. Secondly, Apache Spark is applied for market basket analysis through the preferred Apriori algorithm. Representation of values is accomplished through Pyplot which is a module for Python graphics generation. Our choice of these technologies is based on their ability to scale up and on meeting the requirements of massive data processing on distributed systems.

Data Processing Activities:

Data Ingestion: Input raw transaction data gets called in for the Hadoop Distributed File System (HDFS) storage and processing.

Data Processing: Hadoop MapReduce jobs are mostly used for data cleaning, filtering and any preprocessing and generate key value pairs. Following, we use apriori algorithm with Apache Spark to mine frequent itemset. As a result, we discover association rules. The found patterns help to understand customer behavior and product affinities better.

Data Export: Data being subjected to the analysis, frequent itemset as well as the related association rules are moved to the output files or database for further analysis or visualization using different libraries called matplotlib and seaborn.

Ethical Considerations:

Data sourcing and processing ethics, including data privacy, consenting to usage of data gained from data subjects, and complying with regulations such as GDPR, will be taken into consideration, if not well-handled this can result to anonymization or aggregation of sensitive information.

4. Results

FP Growth Result (Few samples obtained from result generated by FP Growth model)

#	items	freq
1	['JUMBO STORAGE BAG SKULLS']	708
2	['SET/3 RED GINGHAM ROSE STORAGE BOX']	489
3	['WOODEN HEART CHRISTMAS SCANDINAVIAN']	540
4	['HANGING HEART ZINC T-LIGHT HOLDER']	609
5	['JUMBO BAG SCANDINAVIAN BLUE PAISLEY']	670
6	['PACK OF 60 SPACEBOY CAKE CASES']	522
7	['LUNCH BAG WOODLAND']	1037
8	['LUNCH BAG WOODLAND', 'LUNCH BAG RED RETROSPOT']	532
9	['HOT WATER BOTTLE KEEP CALM']	790
10	['STRAWBERRY LUNCH BOX WITH CUTLERY']	585
11	['GARDENERS KNEELING PAD KEEP CALM ']	924
12	['LUNCH BAG BLACK SKULL.']	1295
13	['LUNCH BAG BLACK SKULL.', 'LUNCH BAG RED RETROSPOT']	648
14	['ROUND SNACK BOXES SET OF4 WOODLAND ']	749
15	['COLOUR GLASS T-LIGHT HOLDER HANGING']	633
16	['RED TOADSTOOL LED NIGHT LIGHT']	861
17	['RECIPE BOX PANTRY YELLOW DESIGN']	1164
18	['PACK OF 12 LONDON TISSUES ']	515
19	['SET OF 20 VINTAGE CHRISTMAS NAPKINS']	617
20	['TRAVEL CARD WALLET KEEP CALM']	545
21	['PINK FAIRY CAKE CHILDRENS APRON']	497
22	['SPACEBOY BIRTHDAY CARD']	528
23	['DOORMAT HEARTS']	595
24	['HAND OVER THE CHOCOLATE SIGN ']	680
25	['PAPER CHAIN KIT VINTAGE CHRISTMAS']	832
26	['PAPER CHAIN KIT VINTAGE CHRISTMAS', 'PAPER CHAIN KIT 50'S CHRISTMAS ']	555
27	['6 RIBBONS RUSTIC CHARM']	961
28	['LUNCH BAG RED RETROSPOT']	1607
29	['LUNCH BAG RED RETROSPOT', 'JUMBO BAG RED RETROSPOT']	588
30	['RED KITCHEN SCALES']	570
31	['STRAWBERRY CERAMIC TRINKET BOX']	758

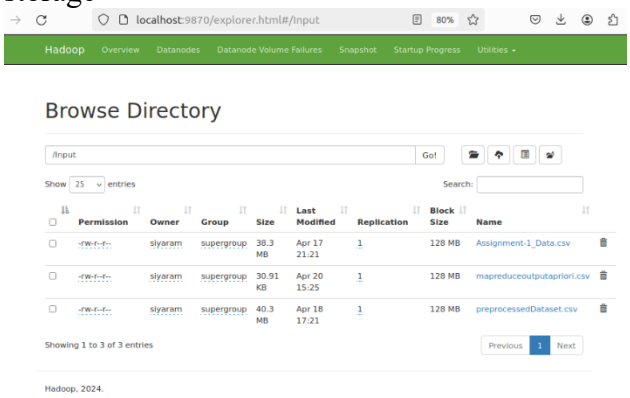
Association Rule (Few samples obtained from result generated by association model)

#	antecedent	consequent	confidence	lift	support
1	['LUNCH BAG SUKI DESIGN ']	['LUNCH BAG BLACK SKULL.']	0.4696282864913871	8.865639165911151	0.021188693090927312
2	['LUNCH BAG SUKI DESIGN ']	['LUNCH BAG RED RETROSPOT']	0.5113327289211242	7.778812211533744	0.023070315376119768
3	['PINK REGENCY TEACUP AND SAUCER', 'GREEN REGENCY TEACUP AND SAUCER']	['ROSES REGENCY TEACUP AND SAUCER ']	0.8524844720496895	18.607757043034606	0.022456743158669776
4	['GARDENERS KNEELING PAD KEEP CALM ']	['GARDENERS KNEELING PAD CUP OF TEA ']	0.5941558441558441	18.987356760886172	0.022456743158669776
5	['JUMBO STORAGE BAG SUKI']	['JUMBO BAG PINK POLKADOT']	0.4313072439633684	8.565530619961295	0.021188693090927312
6	['JUMBO STORAGE BAG SUKI']	['JUMBO BAG RED RETROSPOT']	0.6103247293921732	6.98857548452012	0.029983229028056366
7	['JUMBO STORAGE BAG SUKI']	['JUMBO SHOPPER VINTAGE RED PAISLEY']	0.43713572023313907	9.003080836174854	0.021475027610749785
8	['LUNCH BAG CARS BLUE']	['LUNCH BAG PINK POLKADOT']	0.43222506393861189	9.51946498928596	0.020738740949809792
9	['LUNCH BAG CARS BLUE']	['LUNCH BAG BLACK SKULL.']	0.4484228473998295	8.46532305048929	0.021515932425224645
10	['LUNCH BAG CARS BLUE']	['LUNCH BAG RED RETROSPOT']	0.4782608695652174	7.27569600389599	0.022947600932629 77
11	['WOODEN PICTURE FRAME WHITE FINISH']	['WOODEN FRAME ANTIQUE WHITE ']	0.4874776386404293	12.074332149789843	0.022293123900 68311
12	['GREEN REGENCY TEACUP AND SAUCER']	['ROSES REGENCY TEACUP AND SAUCER ']	0.7417218543046358	16.190066225165562	0.032069374565 38635
13	['GREEN REGENCY TEACUP AND SAUCER']	['REGENCY CAKESTAND 3 TIER']	0.4900662251655629	5.523581838000239	0.021188693090927312
14	['GREEN REGENCY TEACUP AND SAUCER']	['PINK REGENCY TEACUP AND SAUCER']	0.609271523178808	18.595331993947962	0.026342700535 85307
15	['WHITE HANGING HEART T-LIGHT HOLDER']	['RED HANGING HEART T-LIGHT HOLDER']	0.21720243266724587	7.042371182249549	0.020452407248333127
16	['STRAWBERRY CHARLOTTE BAG']	['RED RETROSPOT CHARLOTTE BAG']	0.667574931880109	15.543051771117167	0.020043359103366465
17	['WOODEN FRAME ANTIQUE WHITE ']	['WOODEN PICTURE FRAME WHITE FINISH']	0.5521783181357649	12.074332149789843	0.022293123900 68311
18	['JUMBO BAG BAROQUE BLACK WHITE']	['JUMBO BAG RED RETROSPOT']	0.6261879619852164	7.1702187853 1737	0.024256554996 52309
19	['JUMBO BAG WOODLAND ANIMALS']	['JUMBO BAG RED RETROSPOT']	0.6169724770642202	7.06469608748 8968	0.022006790199206448
20	['LUNCH BAG RED RETROSPOT']	['LUNCH BAG WOODLAND']	0.3310516490354698	7.80445483507245	0.021761361 3122645
21	['LUNCH BAG RED RETROSPOT']	['LUNCH BAG BLACK SKULL.']	0.403235843186061	7.612283133876165	0.026506319793839734
22	['LUNCH BAG RED RETROSPOT']	['JUMBO BAG RED RETROSPOT']	0.3658991910392035	4.189759964091525	0.024052030924 03976
23	['LUNCH BAG RED RETROSPOT']	['LUNCH BAG PINK POLKADOT']	0.3814561294337274	8.40131351014985 1	0.025074651286456417
24	['LUNCH BAG RED RETROSPOT']	['LUNCH BAG CARS BLUE']	0.3490976975731176	7.27569600389599 8	0.022947600932629 77
25	['LUNCH BAG RED RETROSPOT']	['LUNCH BAG SUKI DESIGN ']	0.3509645301804605	7.778812211533742	0.023070315376119768
26	['LUNCH BAG RED RETROSPOT']	['LUNCH BAG APPLE DESIGN']	0.30491599253266954	7.106083192989678	0.020043359103366465
27	['LUNCH BAG RED RETROSPOT']	['LUNCH BAG SPACEBOY DESIGN ']	0.35034225264467955	7.320356453337163	0.023029410561623102
28	['ROSES REGENCY TEACUP AND SAUCER ']	['GREEN REGENCY TEACUP AND SAUCER']	0.7	16.19006622516556	0.032069374565 38635
29	['ROSES REGENCY TEACUP AND SAUCER ']	['REGENCY CAKESTAND 3 TIER']	0.47946428571428573	5.404086395969177	0.021965885384709 78
30	['ROSES REGENCY TEACUP AND SAUCER ']	['PINK REGENCY TEACUP AND SAUCER']	0.5482142857142858	16.731828517924026	0.025115556100953083
31	['GARDENERS KNEELING PAD CUP OF TEA ']	['GARDENERS KNEELING PAD KEEP CALM ']	0.7176470588235294	18.987356760886176	0.022456743158669776

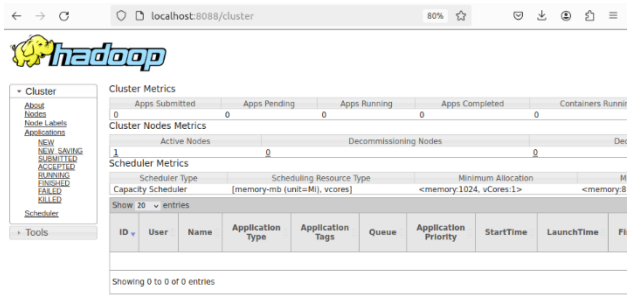
Hadoop Running on Ubuntu:

Initialization of Hadoop with all data nodes and name nodes with resource manager which known as yarn manager

To run Hadoop on localhost, we can use default port 9870, where user can browse directories with can check user logs and configuration of storage

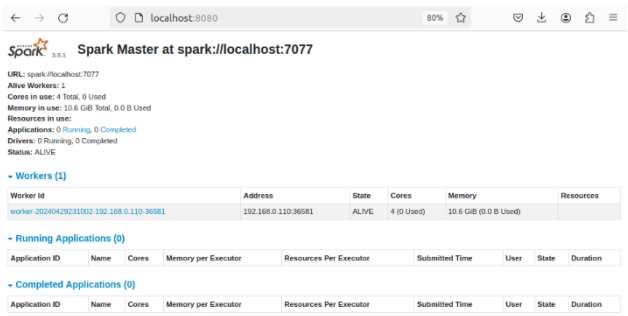


Yarn Manager:



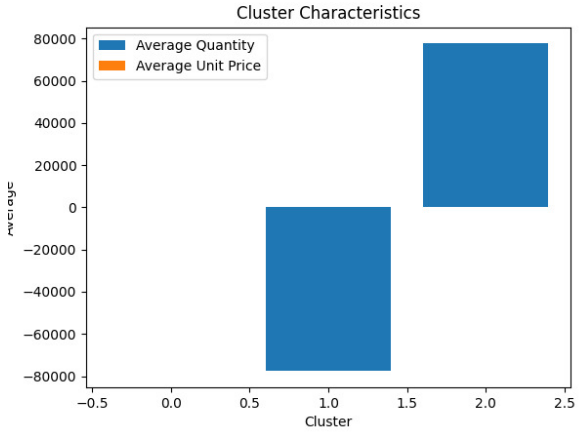
Spark Running on Ubuntu:

To launch spark is the similar process like hadoop, just add variables in environment like java and other path of spark then run in sbin directory to start master node and worker for spark and then submit app for spark to run on cluster like Hadoop with port 8080 on localhost.



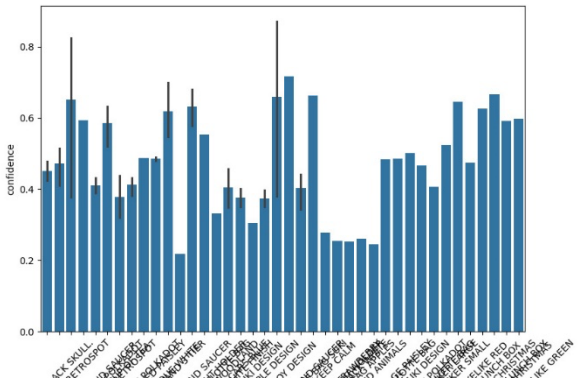
Customer Segmentation Analysis:

We dedicate our first research area in study of the customer segmentation, and we strive to know the unique needs and desires of a retail audience. Through application of sophisticated clustering methods like k-means or hierarchical clustering, we look forward to positioning the customers in specific sections that are based upon their purchasing profiles, demographics, and psychographic attributes. The analysis revealed that the data promise to provide hands-on information that may be applied in adjusting marketing, product, and experience strategies and communication that produce desirable responses from different customer groups.



Market Basket Analysis:

Through market basket analysis - this is the second research product of ours – a fundamental technique in retail analytics, more precisely in shopping analytics, data is used to identify the relationships and patterns between different items that has been gathered from the purchases of final users. Using the world known Apriori algorithm, we aim at discovering the frequent item sets and eventually decoding the actionable association rules which are the securities of the co-purchase instants and product affections. If retailers can find out materials that are relevant with their products, position them in the right section, the retailer can cross-sell and promote the product to make the retailer have a sales revenue and customer satisfaction.



Predictive Analytics for Demand Forecasting:

The third research initiative covers the realm of predictive analytics focusing such forecasting as in demand in retail. By relying on historical sales data, seasonal trends, economic indicators, and other promising factors, we intend to build cutting-edge predictive models which would allow us to accurately forecast the preference of customers concerning different goods. With the ability to have accurate forecasts of demand, retailers can improve the efficiency of their inventory handling processes, reduce disruptions caused by stockouts and overstock situations and better coordinate all supply chain activities while this allows them to increase the profit and the level of customer service.

Commands used for running Hadoop and spark

To run Hadoop:

Just run mrjob file in same directory:

Python3 filename -r Hadoop input -o output

For directory creation on localhost:

hdfs dfs mkdir Input

To move file from local directory to Hadoop directory

Hdfs dfs -put file directory /move directory

For spark:

Spark submit master localhost:7000 filename.py

Ethical Issues

Data Utilization Protocol: It should be a mandatory step to carefully review and follow the Kaggle Terms of Use and the end user license agreement of the dataset. Setting theory for the responsible use of data, redistribution and fair reviews should always be implemented in these agreements. Whenever you are using any dataset always make sure you credit the sources properly and follow the instructions for usage outlined in the terms of conditions.

Data Protection Measures: Prioritize the protection of the data to defend from access by unauthorized person, breaches, misuse. or other threats. Implement robust encryption methods and access control mechanisms to secure the data transfer channels as well as their confidentiality and integrity. Enhancing the level of security, we can restrict the possibilities of data loss and outflows of classified information, which ensures protection of sensitive data.

4. Conclusion

In summing up the Hadoop, Spark, Apriori and our findings, let us get a bit retrospective and express its importance to businesses. Our history is the story of how technology guides us to make smarter actions creating from customer behavior. Firstly, we focused on Hadoop, which is a major part of data processing because of its ability to handle massive amounts of data. Imagine it to be a very big crate where all your data is kept. By means of Hadoop we went through all the huge pile of records fasted and the most efficient way. Subsequently, we introduced Spark procession that could boast of numbers' top speed crunching. There is just performing the calculations, but it is also about comprehension. There the Apriori algorithm showed up. Data analytics function as a detective that uncovers and makes known the relationships between products that are generally used to identify what customers they buy and why. Emphasizing the usage of visualization also plays a crucial role in my learning strategy. By Matplotlib, we could easily transform our data into figures that could be more comprehensible. Through such visuals, businesses will be able to navigate the paths in search of the most lucrative avenues. To conclude, our excursion to see the power of technology and data in business enterprises gave us an insight into the customers behavior, the decision-making processes and customer success in the competitive world of retail.

6. Future Work

Enhanced Personalization Techniques:

To that customer segmentation being very useful, the future investigation should be concentrated on advanced personalization of the product is to be conform each person customer's features. One of these features might be including the sophisticated machine learning algorithms such as collaborative filtering or deep learning to showcase performance-based product suggestions and marketing content is based on individual user tastes and their search history.

Dynamic Pricing Optimization:

Exploring the application of dynamic pricing approach which is real-time based and effects changing their interests of a customers is an attractive standpoint for future research. By the means of predictions algorithms and optimizations algorithms integration it will be possible to adapt price dynamically based on factors including fluctuation of demand, competitors pricing strategies as well as customers segmentation which will not just bring higher revenue but also contribute to high level of competitiveness.

7. References

[1] Li H, Wang Y, Zhang D, Zhang M, Chang EY. Pfp: parallel fp-growth for query recommendation. In: Proceedings of the 2008 ACM conference on recommender systems. RecSys '08. New York: ACM; 2008. p. 107–14. https ://doi. org/10.1145/14540 08.14540 27.

[2.] Zhou L, Zhong Z, Chang J, Li J, Huang JZ, Feng S. Balanced parallel fp-growth with MapReduce. In: 2010 IEEE youth conference on information, computing and telecommunications. 2010. p. 243–6. https ://doi.org/10.1109/YCICT .2010.57130 90

[3] Khan L, Awad M, Thuraisingham B. A new intrusion detection system using support vector machines and hierarchical clustering. VLDB J. 2007;16(4):507–21.https://doi.org/10.1007/s0077 8-006-00025.

[4] Amsterdamer Y, Grossman Y, Milo T, Senellart P. Crowdminer: mining association rules from the crowd. Proc VLDB Endow.2013;6(12):12503.https://doi.org/10.1477 8 /25362 74.25362 88.

[5] Amsterdamer Y, Grossman Y, Milo T, Senellart P. Crowd mining. In: Proceedings of the 2013 ACM SIGMOD international conference on management of data. SIGMOD ’13. New York: ACM; 2013. p. 241–52. https://doi.org/10.1145/24636 76.24653 18.

[6] Naulaerts S, Meysman P, Bittremieux W, Vu TN, Vanden Berghe W, Goethals B, Laukens K. A primer to frequent item-set mining for bioinformatics. Brief Bioinform. 2015;16(2):216.https://doi.org/10.1093/bib/bbt0 7 4.

[7]. Lin X. Mr-apriori: association rules algorithm based on MapReduce. In: Proceedings of IEEE 5th international conference on software engineering and service science. 2014. p. 141–4. https ://doi.org/10.1109/ICSES S.2014.69335 31.

[8]. Barkhordari M, Niamanesh M. Scadibino: an effective MapReduce-based association rule mining method. In: Proceedings of the sixteenth international conference on electronic commerce. ICEC ’14. New York: ACM; 2014. p.
1–118. https ://doi.org/10.1145/26178 48.26178 53.

[9]. Singh S, Garg R, Mishra P. Performance analysis of apriori algorithm with different data structures on Hadoop cluster.
2015. arXiv preprint arXiv :1511.07017