iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 3 **Issue:** V **Month of publication:** May 2015

DOI:

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Association Rule Mining of Intraday Stock Trading Using HADOOP and WEKA

Ms Rajeshwari. K. Gundla (PG Scholar)[1], Prof Mr R.V. Argiddi (Associate Professor)[2]
[1,2]*Department of Computer Science and Engineering, Walchand Institute of Technology, Solapur. Solapur University, MH, India.*

Abstract--Many till date, Stock based association rule mining systems only focused on Interday association rule mining. But in this paper we propose and implement a system which analyses Intraday stock trading sessions and models it into an item transaction datasets using Hadoop based Map-Reduce framework. We thereby facilitate easy association rule mining using Weka classes on large intraday Stock trading session's data.

I. INTRODUCTION

Stock markets globally generate very large datasets. These datasets assist in predicting complex and dynamic patterns and analytical insights with data mining tools. Stock market is a central place for trading stocks. The motivation for our system lies in developing a Hadoop based framework to analyse the large stock datasets efficiently. Analysis of Stock market data involves many attributes, far more than traders can readily and simply understand and interpret them. Traders nonetheless attempt to determine relationships between the data attributes that can lead to profitable trading of financial instruments. Traders apply two types of complementary analysis to trade on the stock market: technical and fundamental analysis. Fundamental analysis: traders study the underlying factors that determine the price of a financial instrument. For example, factors such as a company's profit, market sector, or potential growth can influence the share price. Traders consider these factors more crucial than global concerns such as the general economic trend. Traders have traditionally used technical analysis to trade the market for making profits. Technical analysis is "the study of behaviour of market participants, as reflected in price, volume, and open interest for a financial market, in order to identify stages in the development of price trends." Underlying factors that determine price are ignored in technical analysis, they assume that the price of a financial instrument already quantifies these underlying factors.Technical analysis relies on patterns found directly in the stock data. Because this work relies on the user's finding patterns directly in the data, it is based on technical analysis.
Many traditional analysts don't support the assumptions made by technical analysts, because it ignores the underlying market factors on which stock prices are based and so is thought to be less reliable [14]. In data mining, association rule also known as correlation mining using support and confidence factors is a popular and well researched method for discovering interesting relations between variables in large databases. Piatetsky-Shapiro describes analysing and presenting strong rules discovered in databases using different measures of interestingness [1]. Based on the concept of strong rules, Agrawal et al. introduced association rules for discovering regularities between products in large scale transaction data recorded by point-of-sale (POS) systems in supermarkets [2]. Association rule mining finds interesting associations and/or correlation relationships among large set of data items. Association rules shows attributed value conditions that occur frequently together in a given dataset. Mining association rules on large datasets has received considerable attention in recent years. Association rules are useful for determining correlations between attributes of a relation and have applications in marketing, financial, and retail sectors. Furthermore, optimized association rules are an effective way to focus on the most interesting characteristics involving certain attributes. Optimized association rules are permitted to contain instantiated attributes and the problem is to determine instantiations such that either the support or confidence of the rule is maximized. For example, data are collected using bar-code scanners in supermarkets. Such „market basket" databases consist of a large number of transaction records. Each record lists all items bought by a customer on a single purchase transaction. Managers could use this data for adjusting store layouts, cross-selling, promotions, and catalog design and to identify customer segments based on buying patterns.

II. BACKGROUND

Stock market is a place where the companies and stockholders get revenue. People are trading in the company and it is regular source of income. There are plenty of sources people get the information to make the investment in stock market such as Books,

Internet, news and also from previous experience.

Many times market is in unpredictable conditions so proper investment in stock market is problem for investor. So this approach is suitable for prediction in different sectors of the stock market. Stock market allows companies for publicly trade of the business, or raises the capital with selling the shares of company. It provides companies with access to capital and also for investors with a slice of ownership in the company get the profit based on company's future performance. There are different types of transaction based on trading sessions in the stock market intraday transaction-This is transaction within the day. The term intraday is used to describe the trade on markets during regular business hours, such as opposed stocks and ETFs. This is also called as short term investment. Interday transaction-This is transaction within week or month. Here the investment is for long term.

III. RELATED WORK

News CATS a news based stock trend prediction system was proposed, which took news items as an indicator to predict trends in stock trading scenarios [4]. Several other researchers like [12] used twitter tweets summarization and overall mood as stock market trend predictor. [5] Was one of the foremost work carried out for inter transaction association rule mining laying foundations for association rule mining in stock markets in broader sense. To support multidimensional association rules, Goil et al. [8] presented a scalable parallel system with the techniques of OLAP and data mining to calculate support and confidence. Moreover, Nestorov et al. proposed an approach that can keep all query- processing within the data warehouse and extend association rules using the non-item dimension to obtain more detailed rules. With regards to improving the quality of discovered knowledge, several approaches have been mentioned in the introduction. We also note other interesting research that discusses the similarity between patterns to discover really useful patterns [9] As stated in [2] discovering association rules is an important data mining problem, and there has been considerable research on using association rules in the field of data mining problems. The associations‟ rules algorithm is used mainly to determine the relationships between items or features that occur synchronously in the database. For instance, if people who buy item X also buy item Y, there is a relationship between item X and item Y, and this information is useful for decision makers. Therefore, the main purpose of implementing the association rules algorithm is to find synchronous relationships by analysing the random data and to use these relationships as a reference during decision making [2]. One of the most important problems in modern finance is finding efficient ways to summarize and visualize the stock market data to give individuals or institutions useful information about the market behaviour for investment decisions. The enormous amount of valuable data generated by the stock market has attracted researchers to explore this problem domain using different methodologies. [3] Investigated stock market investment issues on Taiwan stock market using a two stage data mining approach. The first stage Apriori algorithm is a methodology of association rules, which is implemented to mine knowledge and illustrate knowledge patterns and rules in order to propose

stock category association and possible stock category investment collections. Then the K-means algorithm is a methodology of cluster analysis implemented to explore the stock cluster in order to mine stock category clusters for investment information. By doing so, they propose several possible Taiwan stock market portfolio alternatives under different circumstances [3].[13] focuses on employing SVM for identifying evolving trends in stock trading but the approach becomes far complex and near unfeasible for real world implementation . The most notable work to be cited was [14] where authors tried computer vision based but fully dependent on human perception of Stock trading trends using auditory and visual receptors but nevertheless does not address the issue of orchestrating the process. A Recent paper [15] focused on K-means clustering of stock trend patterns and then extracting patterns from clusters using AprioriAll. Hence we propose [16] and implement Asystem for scalable analytics of Intraday Stock Market Trading datasets produced out of heavy intraday transactions, using Hadoop and then extracting Associative patterns among intraday transactions.

IV. PROPOSED SYSTEM AND METHODOLOGY

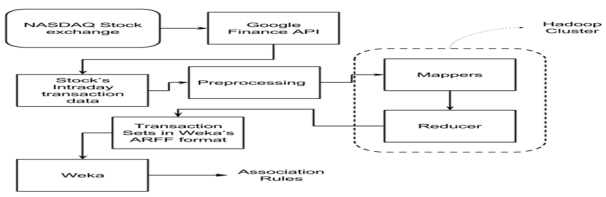


Fig 1: The architecture of proposed system.

Typically the downloaded stock data of any Stock is in following format:
EXCHANGE%3DNASDAQ
MARKET_OPEN_MINUTE=570
MARKET_CLOSE_MINUTE=960
INTERVAL=86400
COLUMNS=DATE,CLOSE,HIGH,LOW,OPEN,VOLUME,CDAYS
DATA=
TIMEZONE_OFFSET=-240
a1092945600,100.335,104.06,95.96,100.01,22353092,1<-- "a"denotes a unix time stamp.
1,108.31,109.08,100.5,101.48,11429498,1<-- The first column (1) * interval (1 day) + last full unix time stamp = date
4,109.4,113.48,109.05,110.76,9140244,1
5,104.87,111.6,103.57,111.24,7632224,1
6,106,108,103.88,104.96,4599110,1
7,107.91,107.95,104.66,104.95,3551168,1
8,106.15,108.62,105.69,108.1,3108977,1
11,102.01,105.49,102.01,105.49,2601620,1<-- Note: We jumped from 8 to 11. We skipped 2 days. This was a weekend.
12,102.37,103.71,102.16,102.32,2463427,1
By subtracting previous close value from immediate subsequent close value an indication of increase or decrease in stock can be inferred, hence the transaction. The transaction set for all stocks under consideration is obtained and are combined to form an overall transaction set table for all stocks, so as to mimic a transaction in conventional market basket analysis.
Since the process outlined above is of parallel nature, it has been implemented using Hadoop's Map-Reduce framework.

We considered following stocks for our sampling. APPL, CSCO, GOOG, MSFT, YHOO.
The overall transaction set obtained after running Hadoop job on the data sets of all companies looks as below.
1,1,1,1,1
1,0,0,0,1
0,1,1,1,1
0,0,1,1,0
0,0,0,1,0
1,1,1,1,1
.
.
.
1,1,0,1,1
1 indicates increase in stocks values as compared to its value in previous minute. 0 indicates drop in stock value or no change. It is to be noted that a row contains data for stock arranged in alphabetic order. If we consider second row in above sample i.e. "1,0,0,0,1" first value 1 corresponds to APPL , second value to CSCO, third value to GOOG , fourth value to MSFT and final value to YHOO.

The combined transaction set generated is prefixed with following Weka's ARFF format header
@RELATION stock
@ATTRIBUTE A {0,1}
@ATTRIBUTE C {0,1}
@ATTRIBUTE G {0,1}
@ATTRIBUTE M {0,1}
@ATTRIBUTE Y {0,1}
@DATA
Subsequently the ARFF formatted data is input to Java program using Weka classes or Weka Tool which mines the files for association rules of positive and negative nature. The association rules obtained indicate incremental or decremental association

among stock during intraday sessions of Stock market. These intraday associations found a vital base for further analysis of Intraday and Historical Stock transactions. The system adopts following step wise process which is automated using various Linuxcron jobs and shell scripts in Linux OS.

Identify the stocks within which the associative patterns have to be identified by their Stock symbols in the Stock market.

Download the identified stocks intraday transactions data periodically using CURL utility in Linux OS and the data have to be saved in files named as that of stock symbols used in Stock Market. E.g. Apples stock's data will be saved in file APPL since Apple's stocks are identified with symbol APPL in NASDAQ stock exchange.

Process the downloaded data periodically to remove the CSV headers

The downloaded files are uploaded to Hadoop cluster's HDFS file system.

Run the Mapper and Reducers written specifically to process the Stocks intraday data and generate transaction item sets for Association rule mining.

Obtain Association rules by running Weka's Association algorithms classes on transaction item sets identified in Step 5.

V. SUMMARY

We considered the Intraday session data obtained per minute during a period of 10 working days earlier to 24Oct2014. On processing the data and analysing it as per the workflow described in the paper we found following association rules.

Instances: 3910, Minimum support: 0.3 (1173 instances), Minimum metric <confidence>: 0.6

Rules found are:

1. A=0 ==> C=0 conf:(0.75)
2. M=0 ==> C=0 conf:(0.74)
3. G=0 ==> C=0 conf:(0.73)
4. Y=0 ==> C=0 conf:(0.72)
5. G=0 ==> M=0 conf:(0.71)
6. C=1 ==> A=1 conf:(0.7)
7. A=0 ==> M=0 conf:(0.7)
8. A=0 ==> Y=0 conf:(0.7)
9. C=0 ==> M=0 conf:(0.69)
10. G=0 ==> Y=0 conf:(0.69)
11. Y=0 ==> M=0 conf:(0.69)
12. M=0 ==> Y=0 conf:(0.68)
13. M=1 ==> G=1 conf:(0.68)
14. Y=0 ==> A=0 conf:(0.68)
15. M=0 ==> G=0 conf:(0.67)
16. C=0 ==> A=0 conf:(0.67)
17. M=1 ==> A=1 conf:(0.67)
18. Y=1 ==> A=1 conf:(0.67)
19. M=0 ==> A=0 conf:(0.67)
20. Y=1 ==> G=1 conf:(0.67)
21. Y=0 ==> G=0 conf:(0.67)
22. C=0 ==> Y=0 conf:(0.67)
23. G=0 ==> A=0 conf:(0.66)
24. A=0 ==> G=0 conf:(0.66)
25. C=0 ==> G=0 conf:(0.65)
26. M=1 ==> Y=1 conf:(0.65)
27. A=1 ==> G=1 conf:(0.65)
28. A=1 ==> Y=1 conf:(0.65)
29. G=1 ==> A=1 conf:(0.65)
30. G=1 ==> M=1 conf:(0.64)

The most interesting rules identify an association that a drop in Apple's stock value also indicates drop in Cisco's stock value as well but an increase in Apple's stock value does not necessarily strongly indicate increase in Cisco's Stock value. These insights could be very well incorporated into Stock portfolio selection. Also in system where news sentiment analysis based predictions are given to improve and weigh effect of news on other stocks. We later verified the patterns observed in real time in month of Nov 2014, which indicated that our approach is a very promising one with good reliability in terms of rule's strength to predict a trend.

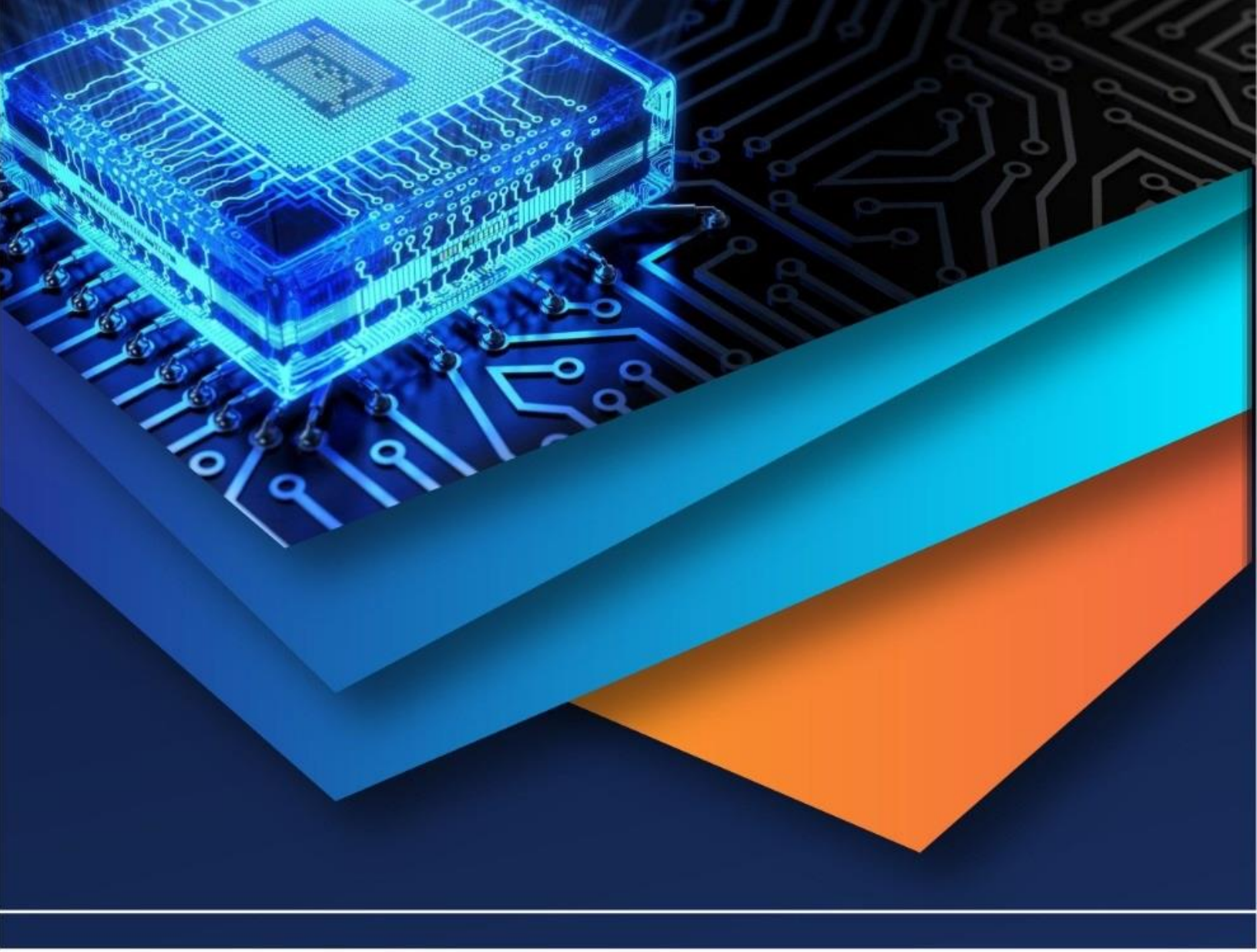
VI. FUTURE WORK

This work can be further supplemented with in database analytics and advanced on the top of Hadoop technologies like Hive and

Pig for SQL like querying the Data.

REFERENCES

[1] Discovery, analysis and presentation of strong rules, Knowledge Discovery in Databases, MIT Press, Cambridge.
[2] Agrawal R, Imilienski T, Swami A (1993). Mining association rules between sets of items in large databases, In Proceedings of the ACM SIGMOD international conference on management of data.
[3] Shu-Hsien L, Hsu-hui H, Hui-wen L (2008). Mining stock category association and cluster on Taiwan stock market, Expert Systems with Applications. 500 Index futures prices, J. Bus. Res.
[4] Mittermayer, M-A., and Gerhard F. Knolmayer. "Newscasts: A news categorization and trading system." Data Mining, 2006.ICDM 06.Sixth International Conference on.IEEE, 2006.
[5] Hongjun Lu, Ling Feng, and Jiawei Han. 2000. Beyond inter transactional association analysis: mining multi dimensionalinter transaction association rules. ACM Trans. Inf. Syst. 18, 4 (October 2000), 423-454. DOI=10.1145/358108.358114
http://doi.acm.org/10.1145/358108.358114
[6] Ke, Yiping, James Cheng, and Wilfred Ng. "Correlated pattern mining in quantitative databases." ACM Transactions on Database Systems (TODS) 33.3 (2008): 14.
[7] Hellerstein, Joseph M., ChristoperRé, Florian Schoppmann, Daisy Zhe Wang, Eugene Fratkin, AleksanderGorajek, KeeSiong Ng et al. "The MADlib analytics library: or MAD skills, the SQL." Proceedings of the VLDB Endowment 5, no. 12 (2012): 1700-1711.
[8] S. Goil and A. Choudhary (1999).A parallel scalable infrastructure for OLAP and data mining.IDEAS, Canada.
[9] S. Tsumoto and S. Hirano (2003).Visualization of rule's similarity using multidimensional scaling, ICDM, USA.
[10] Ke, Yiping, James Cheng, and Wilfred Ng. "Correlated pattern mining in quantitative databases." ACM Transactions on Database Systems (TODS) 33.3 (2008): 14.
[11] Woo,Jongwook, and YuhangXu. "Market basket analysis algorithm with Map/Reduce of cloud computing." The 2011 International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA 2011), Las Vegas. 2011.
[12] Johan Bollen, Huina Mao, "Twitter Mood as a Stock Market Predictor," Computer, vol. 44, no. 10, pp. 91-94, Oct. 2011, doi:10.1109/MC.2011.323
[13] Lean Yu; Huanhuan Chen; Shouyang Wang; Kin Keung Lai, "Evolving Least Squares Support Vector Machines for Stock Market Trend Mining," Evolutionary Computation, IEEE Transactions on , vol.13, no.1, pp.87,102, Feb. 2009.
[14] Nesbitt, Keith V., and Stephen Barrass. "Finding trading patterns in stock market data." Computer Graphics and Applications, IEEE 24.5 (2004): 45- 55.
[15]Wu, Kuo-Ping, Yung-Piao Wu, and Hahn-Ming Lee. "Stock Trend Prediction by Using K-Means and AprioriAll Algorithm for Sequential Chart Pattern Mining."Journal of Information Science and Engineering 30.3 (2014): 653-667.
[16]Rajeshwari.K.Gundla and R.V. Argiddi. "A Framework for Correlated Pattern Mining in Intraday Stock Trading Transactions." International Journal for Scientific Research and Development 2.9 (2014): 713-716.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY