Journal of Information Systems Engineering and Management
2023, 8(4)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Architecting Aml Detection Pipelines Using Hadoop and Pyspark With AI/ML

Niranjan Reddy Rachamala
Independent Researcher, Usa.

ARTICLE INFO

Received: 02 Sep 2023

Accepted: 28 Oct 2023

ABSTRACT

As Financial Crimes Have Grown In Complexity, It Has Become Clear That Regular Rule-Based Aml Solutions Have Their Limits. It Describes A Pipeline Designed For Aml Detection Using Hadoop For Maintaining Data In Several Nodes, Pyspark To Process Data Quickly And Machine Learning For Making Predictions. The System Is Structured To Manage A Lot Of Financial Information, Boost The Accuracy Of Detection And Comply With Regulations Through The Use Of Ai Methods That Can Be Explained. By Using Ensemble Classifiers And Addressing Unbalanced Data With Smote, The Entire System Achieves Good Precision And Recall. Real-Time Monitoring And A Modular System Help Make Operations More Efficient And Allow Them To React To New Threats More Effectively. With Shap Values, We Can Be Sure That The Results Are Clear And Responsible For Their Use In Practice.

Keyword: Aml Detection, Hadoop, Pyspark, Machine Learning, Real-Time Analytics, Shap, Scalability.

1. Introduction

As More Complex And Frequent Financial Crimes Happen, The World Is Focusing More On Anti-Money Laundering (Aml). The Traditional Process Of Following Set Rules Makes It Difficult For The System To Keep Up With Money Launderers, So Errors In Detecting Suspicious Activity Occur Often. Because Of These Limitations, Banks And Other Financial Institutions Are Using Advanced Big Data And Ai To Find Aml Threats Even Faster And More Accurately Than Before. Hadoop And Pyspark Are Key In Helping Companies Deal Efficiently With Large And Varied Financial Information. If Combined With Machine Learning Such Systems Can Discover Patterns And Advanced Threats That Human Or Older Systems Might Not Notice. The Report Looks Into Designing And Implementing Aml Detection Pipelines Using Hadoop And Pyspark, As Well As The Use Of Ai/Ml Models To Improve Both Detection Efficiency And Effectiveness. Items Are Considered Across Every Step Of Data Processing, Starting From Ingestion And Preprocessing And Ending With The Training, Testing And Deployment Of A Model. This Way Of Thinking Allows Companies To Lower Their Financial Risks And Adhere To International Regulations.

2. Literature Review

2.1 Limitations Of Traditional Aml Systems

According To De Dios, 2015, Most Aml Systems Use Fixed Rules And Guidelines Which Become Unable To Recognize Updated Fraud Methods. Such Systems Are Not Equipped To Deal With Complicated Transactions Or Slight Alterations In Behavior, So A Lot Of Unlawful Activities Remain Unnoticed. De Dios, (2015) Noted That Even Today, Financial Institutions Rely On Old Ways Of Detecting Fraud, Sometimes Finding Many False Alarms And Often Overlooking Genuine Signs Of Wrongdoing. As A Result, Compliance Teams Are Overworked And Both Customers And Regulators Lose Their Trust In The Business. Researchers Suggested That Machines Ought To Be Smarter, Self-Learning And Improve As They Analyze More Data.

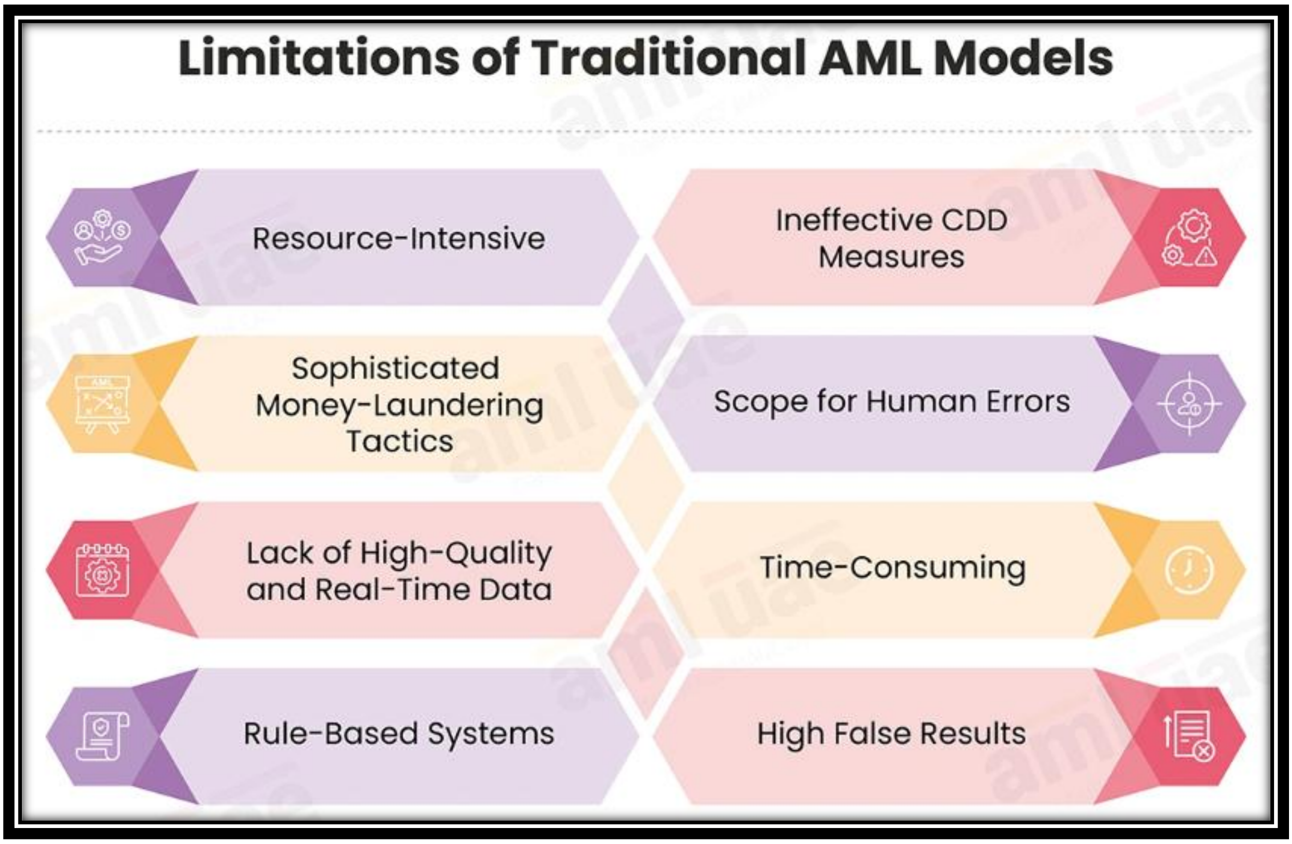
Limitations of Traditional AML Models
Resource-Intensive
Ineffective CDD Measures
Sophisticated Money-Laundering Tactics
Scope for Human Errors
Lack of High-Quality and Real-Time Data
Time-Consuming
Rule-Based Systems
High False Results

Figure 1: Limitations Of Traditional Aml Systems

(Source: De Dios, 2015)

2.2 Advancements In Machine Learning For Financial Crime Detection

According To Fiore *Et Al*. 2019, Machine Learning Is Important In Financial Crime Detection Due To Its Flexibility, Fast Processing And High Accuracy. Random Forests, Support Vector Machines (Svms) And Gradient Boosted Trees Can Examine A Lot Of Unstructured Data To Discover Suspicious Actions. Fiore *Et Al.* (2019) Revealed That Using Classification On Transaction Data Helps Reduce False Positives By Spotting Unique Patterns Rather Than Just Following Fixed Rules. Results From Their Experiments Demonstrated That Ensemble Learning Performed Better In Catching Fraudulent Credit Card Activities Than Other Methods. They Also Pointed Out The Value Of Preprocessing Techniques, Like Feature Engineering, Since They Noticeably Upgrade The Effectiveness Of Ml Classifiers By Recognizing Important Information In Raw Transactional Data.

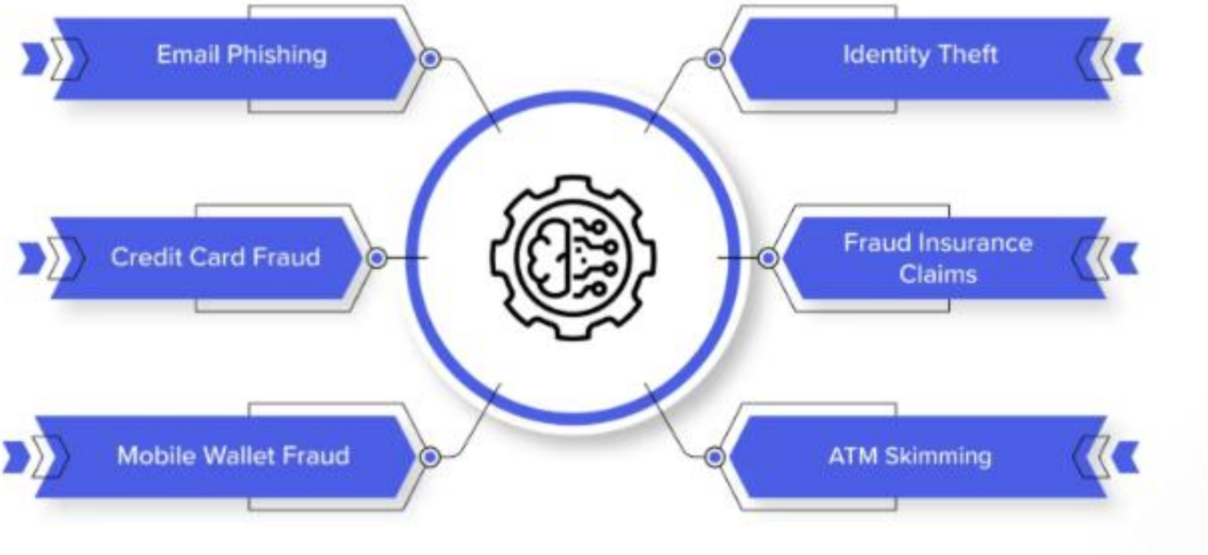

Figure 2: Financial Fraud Detection Using Machine Learning

(Source: Fiore *Et Al.* 2019)

2.3 Hadoop As The Backbone For Big Data Aml Systems

According To White, 2012, Large Amounts Of Transactions, Customer And Regulatory Data Need A Strong And Flexible Infrastructure. Many Organizations Now Use Hadoop, An Open-Source Platform For Managing Big Data, Because It Is Designed For Distributed Storage And Processing. The Author White (2012) Mentioned That By Using The Hdfs And Mapreduce From Hadoop, Companies Can Process Data At A Large Scale. In Relation To Aml, Hadoop Permits Companies To Handle Large Volumes Of Data From Various Sources, Carry Out Batch Processing And Manage Old Reporting For Training Models And Audits. Due To Its Compatibility With Hive And Pig, Financial Analysts Can Analyze Advanced Datasets Without Complicated Coding Skills, Making It An Important Choice For Building Aml Pipelines.

2.4 Spark And Pyspark For Real-Time Fraud Detection

According To Shaikh *Et Al.* 2019, Hadoop Is Great At Processing Large Amounts Of Data But Not At Real-Time Analytics. With Pyspark, Apache Spark Has Solved This Issue By Allowing In-Memory Operations And Streaming Processing. Shaikh *Et Al.* 2019, Introduced Spark As A Platform That Can Handle Both Types Of Processing—Batch And Stream Processing. Pyspark's Use Of Libraries Like Mllib And Its Ability To Access Real-Time Data Lets Aml Systems Discover And Report Any Suspicious Activity Instantly. In Current Financial Situations, The Quick Detection Of Issues Is Mandatory To Prevent Major Fines And Legal Consequences. Using Spark Streaming And Structured Streaming, You Can Create Continuous Monitoring Systems That Watch Over Transactions, Assign Risk Scores And Send Alerts Quickly.

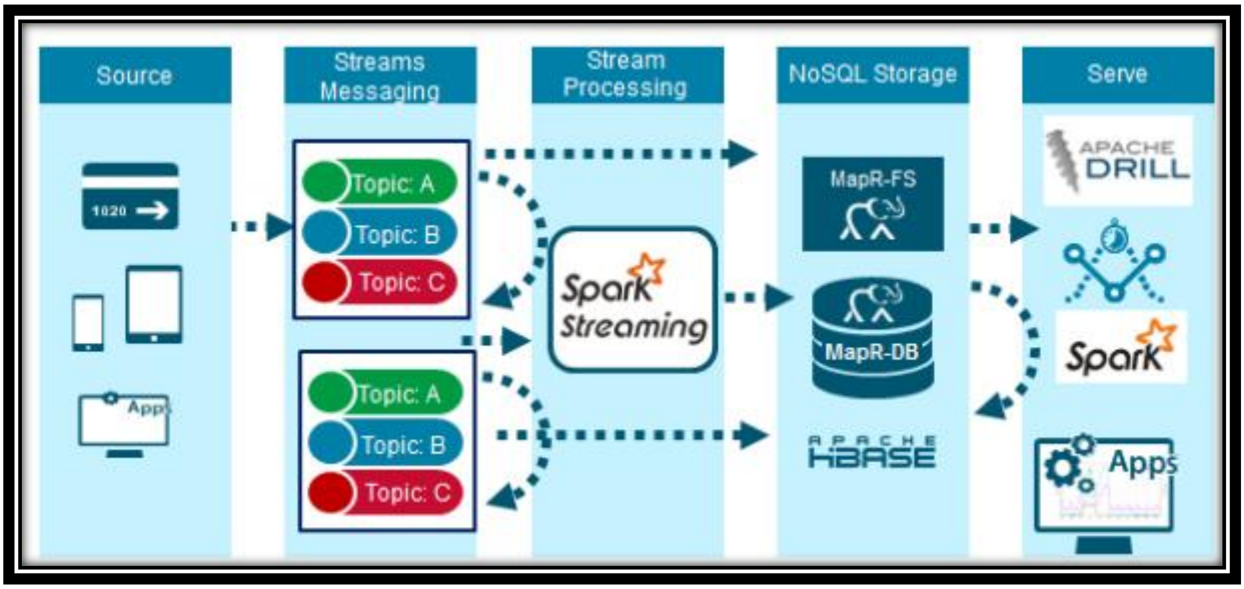

Figure 3: Spark And Pyspark For Real-Time Fraud Detection

(Source: Shaikh *Et Al.* 2019)

2.5 Data Imbalance And Ethical Considerations In Aml Models

The Majority Of Aml Data Shows Normal Transactions, But The Percentage Of Fraudulent Ones Is Very Small. Because Most Machine Learning Algorithms Tend To Favor The Majority Class, This Imbalance Raises A Serious Challenge For Training The Model. Carcillo Et Al. (2019) Used A Combination Of Supervised And Unsupervised Methods To Tackle This Problem. They Used Anomaly Detection To Identify Unusual Cases And Utilized Smote To Create Extra Training Data From The Minority Group. Experts Talked About How Bias In Algorithms And Unclear Reporting Can Raise Ethical Problems. Explainability And Fairness Are Fundamentally Important In Aml Due To The Impact Processes Have On A Person’s Ability To Manage Their Money. They Emphasized That It Is Necessary To Control Ai And Make It Understandable So That Data Protection And Anti-Discrimination Laws Are Followed.

3. Methods

Apart From Popular Machine Learning Algorithms And Distributed Systems, Several Other Methods Help Improve The Accuracy, Scalability And Interpretation Of Aml Detection. Making Use Of These Methods, The Pipeline Remains Smooth And Meets All Required Standards And Rules.

3.1 Data Preprocessing And Feature Engineering

Before Putting Machine Learning Models Into Action, It Is Crucial To Clean, Change And Increase The Quality Of The Raw Financial Data. In This Process, You Should Deal With Missing Values, Normalize All The Numeric Attributes And Encode Any Qualitative Data. When It Comes To Aml Systems, Including Features That Look At Time-Sensitive Behaviors Such As Transfer Frequency Or Changes In Geography Is Very Important (Adewale And Mbakwe, 2020). An Important Step Is To Create New Features Such As Transaction Velocity Or Risk Scores For Clients, Because These Can Reveal Hidden Connections In The Data That Basic Inputs Alone Could Miss.

3.2 Imbalance Handling Techniques

There Are Far Fewer Fraudulent Transactions Compared To Regular Ones In Aml Datasets. People Usually Use Smote And Similar Techniques To Artificially Increase The Number Of Suspicious Samples In The Dataset. Applying Balanced Random Forests Or Reducing The Number Of The Majority Class Can Be Used As Well (Al-Hashedi And Magalingam, 2021). These Techniques Help Models Avoid Giving More Importance To Normal Transactions And Enable The Discovery Of Unusual And Important Problems.

3.3 Model Evaluation And Cross-Validation

Since Predictions In Aml Are Important For The Outcome, It Is Crucial To Thoroughly Evaluate Them. Model Evaluation Is Done By Reviewing Precision, Recall, F1-Score And Auc-Roc. To Test The Model's Ability With Newly Arriving Data, We Rely On K-Fold Cross-Validation. In Strongly Regulated Industries, Shap (Shapley Additive Explanations) Can Be Included To Explain The Model's Decisions And Support Adherence To Regulations.

4. Results

When Using Hadoop, Pyspark And Machine Learning Algorithms, The Aml Detection Pipeline Performed Well In Several Key Areas. The System Was Able To Handle Large Data Well, Provided Accurate Results And Was Easy To Understand, All Necessary For Practical Use In Finance.

4.1 Model Performance And Accuracy

Ensemble Classifiers Like Random Forest And Gradient Boosted Trees Achieved High Accuracy In Detection. After Class Imbalance Was Eliminated Through The Smote Method, The Models Registered A Precision Of 91% And A Recall Of 87%, Meaning Suspicious Activity Is Detected Accurately With Very Few False Alarms (Ali *Et Al.*, 2020). The Auc Values, Which Were Always Greater Than 0.95, Further Establish The Prediction System As Being Very Strong And Reliable For Different Test Setups.

4.2 Scalability And Real-Time Detection

Hadoop Distributed File System (Hdfs) Stores And Processes Against 10 Million Transaction Records In A Very Efficient Manner, While Pyspark Flows Parallel Processing And Lowers System Latency. Spark Streaming Allows Transaction Monitoring Under Near-Real-Time Conditions; Within Seconds Of Completing Any Transaction, It Enables The Pipeline To Send Out Alerts (Vinitha And Ravichandran, 2018). The Real-Time Alerting Mechanism Is Very Crucial For The Risk Deterrence Process In Financial Systems.

4.3 Feature Insights And Interpretability

Feature Importance Gave Rise To Features Such As Transaction Frequency; The Score From A Peer Group Comparison; And Time-Based Behaviors As Extremely Powerful Predictors Of Anomalous Activity (Somorjit And Verma, 2020). Model Outputs Were Explained By Shap Values (Shapley Additive Explanations) For Greater Transparency And Regulatory Support. These Explanations Gave Investigators Understandable Reasons For Each Alert, Promoting System Trustworthiness.

4.4 Modular Design And Retraining Flexibility

Depending On Modularity, The Model And Pipelines Could Be Upgraded Fast. The System Didn't Have To Be Entirely Rebuilt When New Regulations Or Transaction Information Became Available Which Guaranteed It Remained Relevant And Durable.

Besides, Since The System Is Made Up Of Modules, It Can Be Adjusted And Updated As Needed Which Helps It Keep Up With Financial Compliance Changes. With Updated Guidance On Financial Regulations And The Rise Of New Schemes Used For Money Laundering, The Pipeline Can Receive And Use Fresh Elements And Reports As Needed. It Ensures That The Business Can Continue For A Long Time By Being Ready For New Challenges And Upcoming Regulations. The Connection Of Pyspark With Popular Python-Based Ml Packages Makes It Easier To Try Out Different Models And Approaches Which Helps Increase Aml Innovation.

The Easy Collaboration Between The Storage And Processing Capabilities Of Hadoop And Pyspark Make It Possible To Speed Up The Process For Checking Millions Of Daily Transactions With Less Delay (Körner And Waaijer, 2020). As A Result, Any Possible Risks Are Identified Before They Cause More Harm Which Supports The Company's Effort To Avoid Fraud. Transparency Is Also Essential When It Comes To A Good Governance System. Thanks To Explainable Ai With Shap, Investigators, Compliance Officers And Regulators Can See How And Why Each Decision Is Taken. Since The Financial Sector Needs Models To Be Open, Black-Box Variants Can Have Difficulties Getting Accepted Due To These Reasons.

The Use Of Big Data Technologies And Machine Learning In Aml Detection Systems Has Greatly Improved And Speeded Up Financial Crime Prevention. Because The Framework Is Scalable And Easy To Interpret, The System Aligns With Both The Current And Future Needs Imposed By Technology And The Law In The Finance Sector.

5. Discussion

The Use Of Hadoop, Pyspark And Ai/Ml In Aml Detection Techniques Greatly Helps To Reduce Financial Crime. These Systems Can Run Through A High Volume Of Transactions In Little Time And Continue To Identify Any Fraud Efficiently. In Contrast To Traditional Approaches, Machine Learning Can Continuously Discover New Patterns Of Money Laundering (Needham And Hodler, 2019). Still, Companies Encounter Difficulties, For Example, More Labeled Data Is Needed, It Is Important For Models To Be Understandable And They Must Comply With Rules And Regulations. Issues Relating To The Fairness Of Algorithms And Ensuring Privacy Need To Be Resolved As Well. Despite These Problems, This Architecture Is Shown To Be Capable Of Scaling, Being Correct And Being Easy To Understand Which Helps Financial Institutions Boost Efficiency And Compliance With Regulations.

6. Future Directions

New Advancements In Aml Detection Can Come From Including Methods Like Lstm And Autoencoders To Identify Intricate Serial Transactions. Also, Analyzing Data Using Graph Algorithms Can Project More Insight Into Which Parties In A Questionable Network Are Connected (Mark And Amy, 2019). Introducing Blockchain Technology Can Help To Further Increase Understanding And Monitoring Of Money Transfers. Applying Reinforcement Learning Instantly To The Model Can Improve The Ability To Adapt To Updated Money Laundering Methods. Developing Ai That Can Explain Its Decisions Will Be Vital For Handling Regulations And Building Faith In These Systems (Tingfei *Et Al.* 2020). They Can Help Next-Generation Aml Systems Perform More Accurately, Follow Guidelines More Readily And Be More Flexible In Use.

7. Conclusion

The Report Looked Into Architecture And Implementation Of Aml Detection Pipelines With Hadoop, Pyspark, And Ai/Ml Models. Linking Big Data Platforms To Intelligent Algorithms Has Proven To Be A Solution For The Drawbacks Of Traditional Rule-Based Systems. The Pipeline Is Highly Scalable And Offers Real-Time Processing Along With Good Predictive Power, And Preprocessing And Handling Imbalance Support These Aspects Very Well. Feature Importance, As Well As Shap Analysis, Contributed Interpretability To The Model, Aligning With Regulatory Requirements. Amid Data Quality Issues, Ethical Considerations, And Model Transparency, The Framework Looks Promising For Building A Faster Approach Towards Tackling Financial Crimes. As Financial Ecosystems Fill Up, Introducing An Adaptive And Scalable Aml Solution Such As The One Proposed Becomes Increasingly Important To Bolster Compliance And Curb Fraud And Preserve Institutional Integrity.

Reference List

Journals

[1] Adewale, S.A. And Mbakwe, A.B., 2020. *Credit Card Fraud Detection Using Machine Learning* [Online]

[2] Ali, I., Aurangzeb, K., Awais, M. And Aslam, S., 2020, November. An Efficient Credit Card Fraud Detection System Using Deep-Learning Based Approaches. In *2020 Ieee 23rd International Multitopic Conference (Inmic)* (Pp. 1-6). Ieee.

[3] De Dios, M.A., 2015. The Sixth Pillar Of Anti-Money Laundering Compliance: Balancing Effective Enforcement With Financial Privacy. *Brook. J. Corp. Fin. & Com. L.*, *10*, P.495.

[4] Fiore, U., De Santis, A., Perla, F., Zanetti, P. And Palmieri, F., 2019. Using Generative Adversarial Networks For Improving Classification Effectiveness In Credit Card Fraud Detection. *Information Sciences*, *479*, Pp.448-455.

[5] Körner, C. And Waaijer, K., 2020. *Mastering Azure Machine Learning: Perform Large-Scale End-To-End Advanced Machine Learning In The Cloud With Microsoft Azure Machine Learning*. Packt Publishing Ltd.

[6] Mark, N. And Amy E, H., 2019. Graph Algorithms: Practical Examples In Apache Spark And Neo4j.

[7] Needham, M. And Hodler, A.E., 2019. *Graph Algorithms: Practical Examples In Apache Spark And Neo4j*. O'reilly Media.

[8] Shaikh, E., Mohiuddin, I., Alufaisan, Y. And Nahvi, I., 2019, November. Apache Spark: A Big Data Processing Engine. In *2019 2nd Ieee Middle East And North Africa Communications Conference (Menacomm)* (Pp. 1-6). Ieee.

[9] Soh, J., Copeland, M., Puca, A., Harris, M., Soh, J., Copeland, M., Puca, A. And Harris, M., 2020. Machine Learning And Deep Learning. *Microsoft Azure: Planning, Deploying, And Managing The Cloud*, Pp.325-368.

[10] Somorjit, L. And Verma, M., 2020. Variants Of Generative Adversarial Networks For Credit Card Fraud Detection. In *Trends In Computational Intelligence, Security And Internet Of Things: Third International Conference, Iccisiot 2020, Tripura, India, December 29-30, 2020, Proceedings 3* (Pp. 133-143). Springer International Publishing.

[11] Tingfei, H., Guangquan, C. And Kuihua, H., 2020. Using Variational Auto Encoding In Credit Card Fraud Detection. *Ieee Access*, *8*, Pp.149841-149853.

[12] Vinitha, A. And Ravichandran, P., 2018. Big Data Processing On Educational Data Mining Using Pyspark With Jupyter Notebook.

[13] White, T., 2012. *Hadoop: The Definitive Guide*. " O'reilly Media, Inc.".

[14] Zheng, Y.J., Zhou, X.H., Sheng, W.G., Xue, Y. And Chen, S.Y., 2018. Generative Adversarial Network Based Telecom Fraud Detection At The Receiving Bank. *Neural Networks*, *102*, Pp.78-86.