BIG DATA TECHNOLOGIES (CS 583-C)

FINAL PROJECT

ANTI-MONEY LAUNDERING DETECTION

ARUN KASHYAP
M.S Data Science
Stevens Institute of Technology
Fall 2024

PROBLEM STATEMENT

OBJECTIVE:

- To detect money laundering activities, a multi-billion-dollar issue by analyzing high-volume transactional data.
- Develop a scalable and efficient fraud detection system leveraging distributed big data frameworks.

PROBLEM STATEMENT:

- Money laundering is a multi-billion-dollar global issue, enabling criminal networks and disrupting economies.
- Detection is notoriously difficult as criminals work hard to cover their tracks.
- Challenges in Detection Systems

DATASET

- SOURCE:
  - IBM Transactions for Anti Money Laundering (AML)
  - The dataset (~3GB) contains synthetic financial transactions labeled for anti-money laundering (AML) detection. (31 Million rows)
  - Key features like timestamp, sender, receiver, amount, currency, and is_laundering.
  - **Available in various sizes (700MB to 17GB) and includes a range of transaction amounts.**
  - **Dataset used in this project belongs to Group LI, characterized by a relatively lower ratio of illicit transactions.**

Timestamp	From Bank	Account	To Bank	Account.1	Amount Received	Receiving Currency	Amount Paid	Payment Currency	Payment Format	Is Laundering
9/1/2022 0:29	123878	82E2AA140	123878	82E2AA140	5.08	Euro	5.08	Euro	Reinvestn	0
9/10/2022 22:59	718	8005B38B0	215893	807A48510	136.19	US Dollar	136.19	US Dollar	Cheque	0
9/16/2022 7:45	66015	830E590E0	1146077	8439C5060	860.94	US Dollar	860.94	US Dollar	Cheque	0
9/13/2022 1:40	16932	80392AD10	2597	803A7B870	369.03	US Dollar	369.03	US Dollar	Cheque	0
9/1/2022 12:54	70	10042B660	7049	805DBDFE0	2414.27	US Dollar	2414.27	US Dollar	Credit Car	0
9/14/2022 11:53	2136660	84CB70F00	188829	84DF88D00	1033.48	US Dollar	1033.48	US Dollar	Cheque	0

BIG DATA INFRASTRUCTURE

1. AWS EMR CLUSTER CONFIGURATION:

   - Primary Node: 1 x m5.xlarge (4 vCPUs, 16 GiB RAM, 64 GiB EBS).
   - Core Nodes: 1 x m5.xlarge (4 vCPUs, 16 GiB RAM, 64 GiB EBS).
   - Task Nodes: 2 x m5.xlarge (4 vCPUs, 16 GiB RAM each).

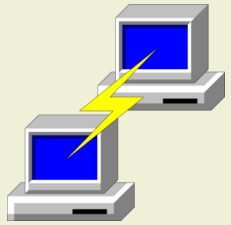
2. STORAGE:

   - Dataset stored in an Amazon S3 bucket for scalable and durable storage.

3. TOOLS AND ENVIRONMENT:

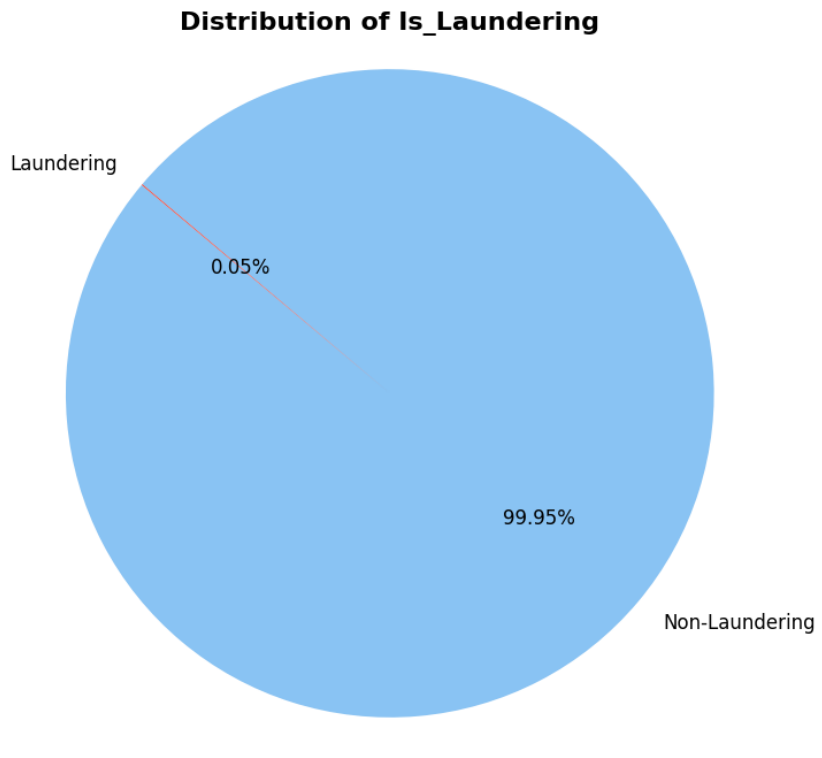
   - PySpark: Used for distributed data processing on EMR.
   - PuTTY: For secure SSH connection to the EMR cluster.
   - JupyterHub: Deployed on the primary node for developing and running PySpark code.

Amazon S3



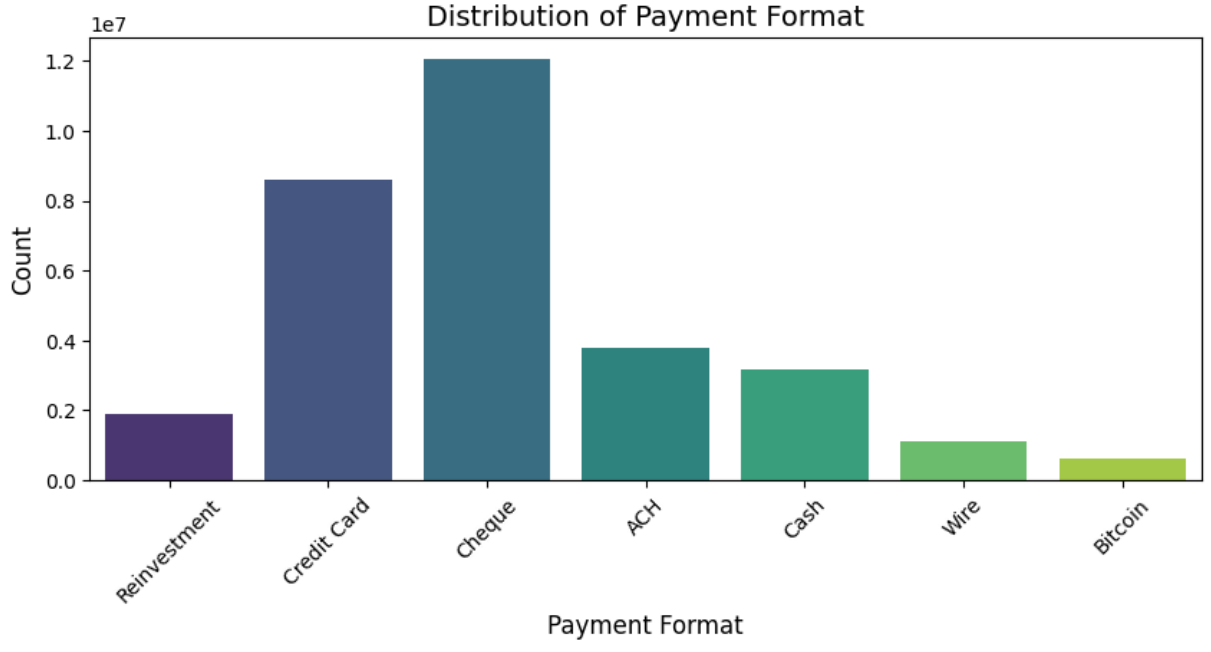
EXPLORATORY DATA ANALYSIS (EDA)

DISTRIBUTION OF Is_Laundering:



- Non-laundering transactions dominate the dataset (99.95%), with only 0.05% labeled as laundering.

DISTRIBUTION OF PAYMENT FORMAT:

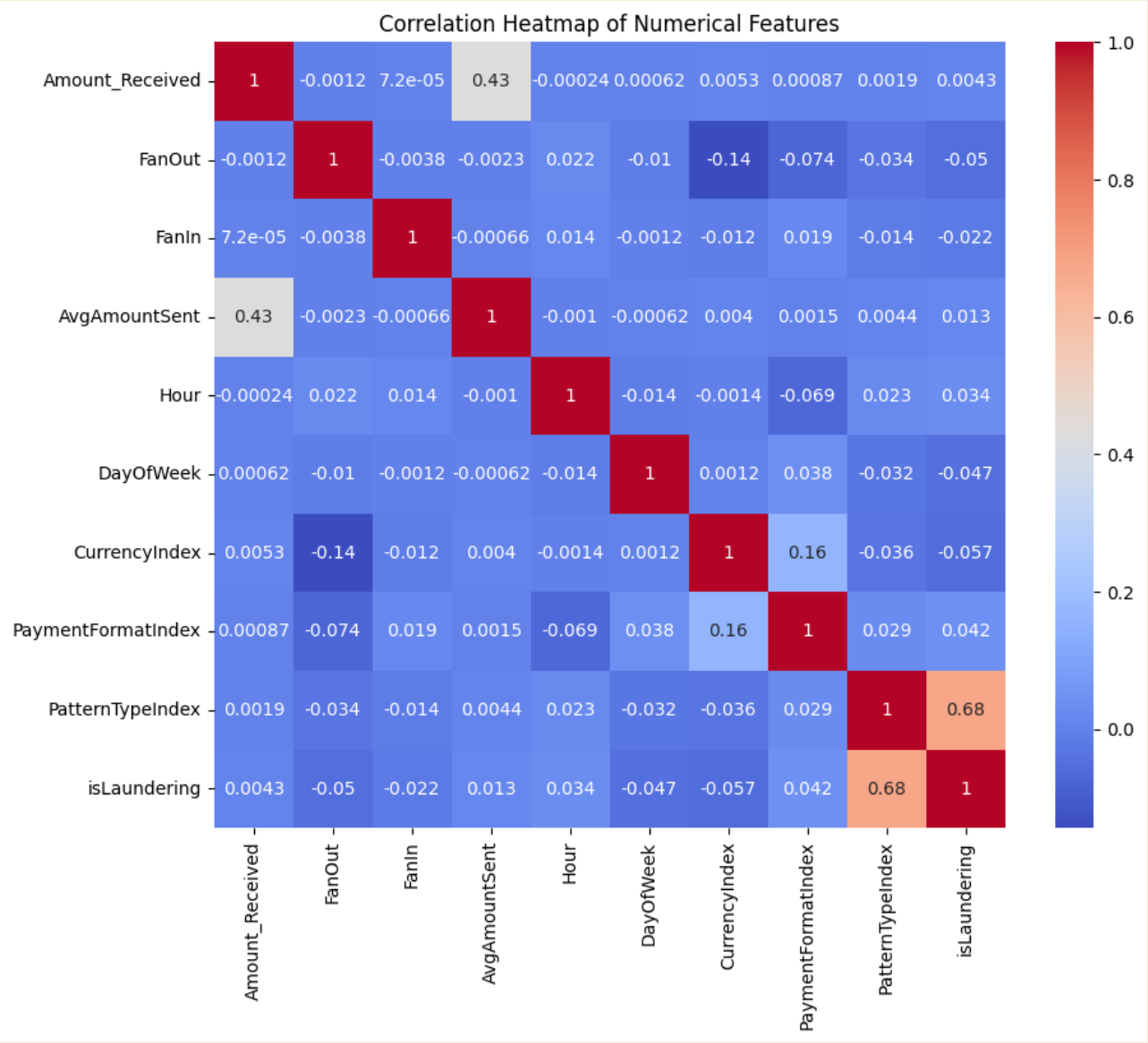


- Cheques and Credit Cards are the most common payment formats, with Bitcoin and Wire transfers being the least frequent.

EXPLORATORY DATA ANALYSIS (EDA)

FEATURE CORRELATION HEATMAP:

- Features like PatternTypeIndex have a notable correlation with isLaundering (0.68), indicating a potential risk of data leakage.
- Other features (FanOut, AvgAmountSent) show weaker correlations with isLaundering.



DATA PREPROCESSING

FEATURE ENGINEERING:

- Extracted temporal features from timestamp (e.g., hour, day, day_of_week).
- Derived features like fan_out, fan_in, and average_amount_sent.

LAUNDERING PATTERN MAPPING:

- A .txt file that outlines laundering patterns such as scatter-gather strategy was mapped to its corresponding laundering pattern (if applicable).

```
root
 |-- Timestamp: string (nullable = true)
 |-- From_Bank: string (nullable = true)
 |-- From_Account: string (nullable = true)
 |-- To_Bank: string (nullable = true)
 |-- To_Account: string (nullable = true)
 |-- Amount_Received: float (nullable = true)
 |-- Receiving_Currency: string (nullable = true)
 |-- Amount_Paid: float (nullable = true)
 |-- Payment_Currency: string (nullable = true)
 |-- Payment_Format: string (nullable = true)
 |-- Pattern_Type: string (nullable = true)
 |-- isLaundering: integer (nullable = true)
```

Schema of the joined dataset

- Grouped the dataset by the *'is_laundering'* column to generate a total count of transactions for laundering vs. non-laundering cases.

```
+------------+--------+
|isLaundering|   count|
+------------+--------+
|           1|    9862|
|           0|31241621|
+------------+--------+
```

DATA PREPROCESSING

HANDLING DATA IMBALANCE:

Steps Taken:

- Dropped 50% of the majority class records to reduce the imbalance.

isLaundering	count
1	9862
0	15618450

- Applied SMOTE (Synthetic Minority Oversampling Technique) to generate synthetic samples for the minority class, increasing its size to 1.2 million records.

isLaundering	count
0	15618450
1	1183440

MODEL SELECTION AND TRAINING

CHOSEN MODEL: RANDOM FOREST

REASON FOR SELECTION:

- Handles imbalanced datasets effectively.
- Provides feature importance insights.
- Performs well on classification tasks with tabular data.

TRAINING CONFIGURATION (FIRST RUN):

- Features: FanOut, FanIn, AvgAmountSent, Hour, DayOfWeek, CurrencyIndex, PaymentFormatIndex, PatternTypeIndex.
- Hyperparameters:
  - Number of Trees: 20.
  - Max Depth: 10.
  - Max Bins: 75.
- Target Variable: isLaundering.

TRAINING PROCESS:

- Dataset split into training (70%), validation (15%), and test (15%) sets.
- Used binary classification evaluator for performance evaluation on the validation set.

FIRST RUN RESULTS

- PERFORMANCE METRICS:
    - F1 Score: 1.0
    - ROC-AUC: 1.0
    - Precision: 1.0
    - Recall: 1.0
- OBSERVATIONS:.
    - Unrealistically high performance: Indicates potential data leakage due to the inclusion of the PatternTypeIndex feature, which may directly encode laundering information.
- INFERENCE:
    - The feature PatternTypeIndex was identified as a cause of data leakage and removed in the subsequent training run.

MODEL TRAINING (SECOND RUN)

- UPDATED CONFIGURATION:
  - Removed Feature: PatternTypeIndex.
  - Features Used: FanOut, FanIn, AvgAmountSent, Hour, DayOfWeek, CurrencyIndex, PaymentFormatIndex.

- TRAINING PROCESS:
  - Same hyperparameters and dataset splits as the first run:
    - Number of Trees: 20.
    - Max Depth: 10.
    - Max Bins: 75.

- Evaluated using binary classification metrics.

SECOND RUN RESULTS

- PERFORMANCE METRICS:
    - F1 Score: 0.90
    - ROC-AUC: 0.95
    - Precision: 0.97
    - Recall: 0.98
- OBSERVATIONS:
    - Metrics reflect realistic and reliable performance after resolving data leakage.
    - Feature selection significantly improved model robustness.

PERFORMANCE COMPARISON

No.	**Metric**	**First Run**	Second Run
1	F1-Score	1.0	0.90
2	ROC-AUC	1.0	0.95
3	Precision	1.0	0.97
4	**Recall**	1.0	0.98

KEY TAKEAWAY:

- Removing *PatternTypeIndex* resolved data leakage, leading to reliable model performance.

FEATURE IMPORTANCE:

- FanOut: Strongest predictor of laundering activity.
- AvgAmountSent: Highlights unusual transaction sizes.
- CurrencyIndex: Indicates laundering trends across currencies.

PERFORMANCE

- PREDICTION CONFUSION MATRIX:

```
# Calculate confusion matrix
final_predictions.crosstab("isLaundering", "prediction").show()
```

```
+----------------------+-------+------+
|isLaundering_prediction|    0.0|   1.0|
+----------------------+-------+------+
|                     0|2337671|  5865|
|                     1|  48794|127961|
+----------------------+-------+------+
```

True Negatives (TN) (2337671):

- Actual class is 0.0 (not laundering) and predicted class is also 0.0.
- The model correctly identified these transactions as legitimate.

False Negatives (FN) (48764):

- Actual class is 1.0 (laundering) but predicted as 0.0 (not laundering).
- These are laundering transactions that were missed by the model.

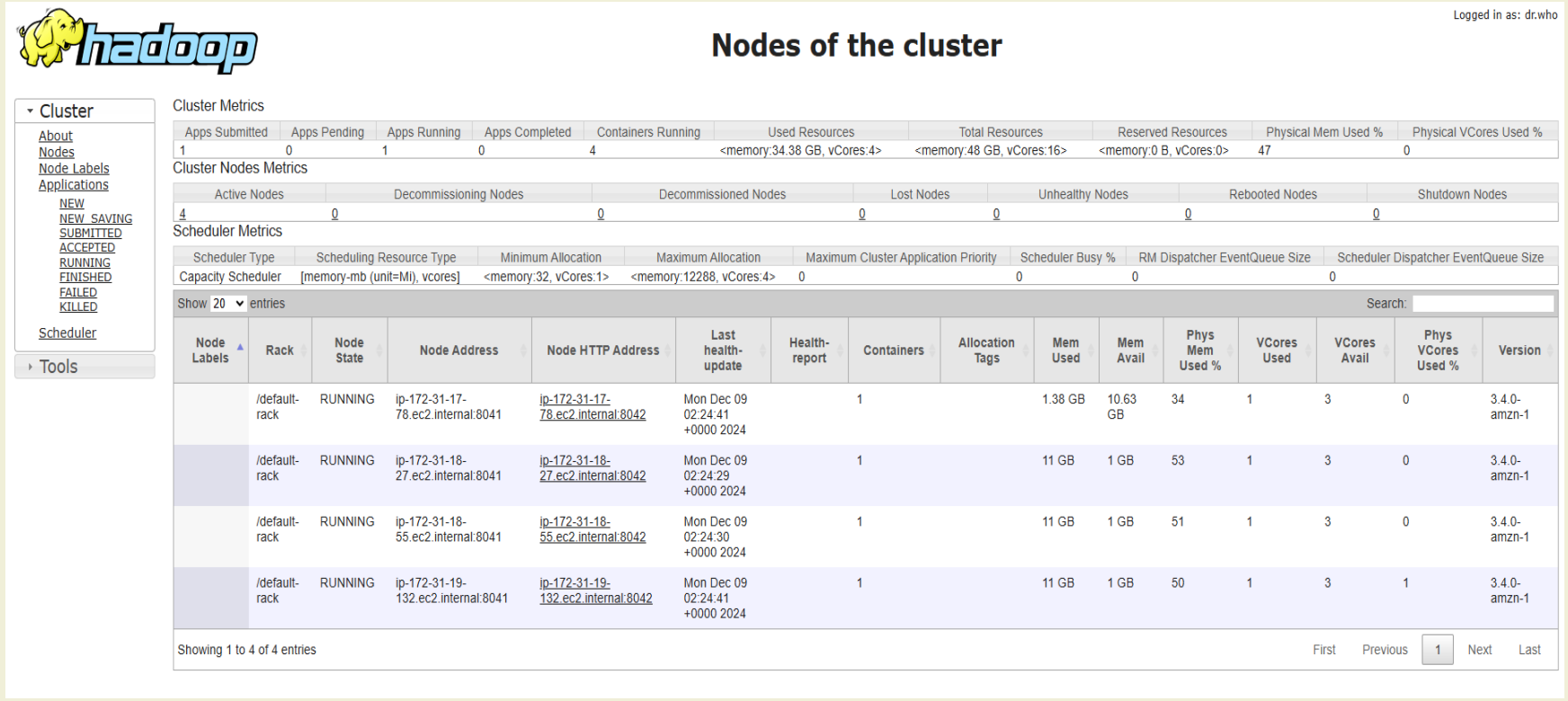
False Positives (FP) (5865):

- Actual class is 0.0 (not laundering) but predicted as 1.0 (laundering).
- These are legitimate transactions that were incorrectly flagged as laundering.

True Positives (TP) (127961):

- Actual class is 1.0 (laundering) and predicted class is also 1.0.
- The model correctly identified these transactions as laundering.

RESOURCE UTILIZATION



Logged in as: dr.who

Nodes of the cluster

- Cluster
  - About
  - Nodes
  - Node Labels
  - Applications
    - NEW
    - NEW_SAVING
    - SUBMITTED
    - ACCEPTED
    - RUNNING
    - FINISHED
    - FAILED
    - KILLED
  - Scheduler
- Tools

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Used Resources	Total Resources	Reserved Resources	Physical Mem Used %	Physical VCores Used %
1	0	1	0	4	<memory:34.38 GB, vCores:4>	<memory:48 GB, vCores:16>	<memory:0 B, vCores:0>	47	0

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes	Rebooted Nodes	Shutdown Nodes
4	0	0	0	0	0	0

Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation	Maximum Cluster Application Priority	Scheduler Busy %	RM Dispatcher EventQueue Size	Scheduler Dispatcher EventQueue Size
Capacity Scheduler	[memory-mb (unit=Mi), vcores]	<memory:32, vCores:1>	<memory:12288, vCores:4>	0	0	0	0

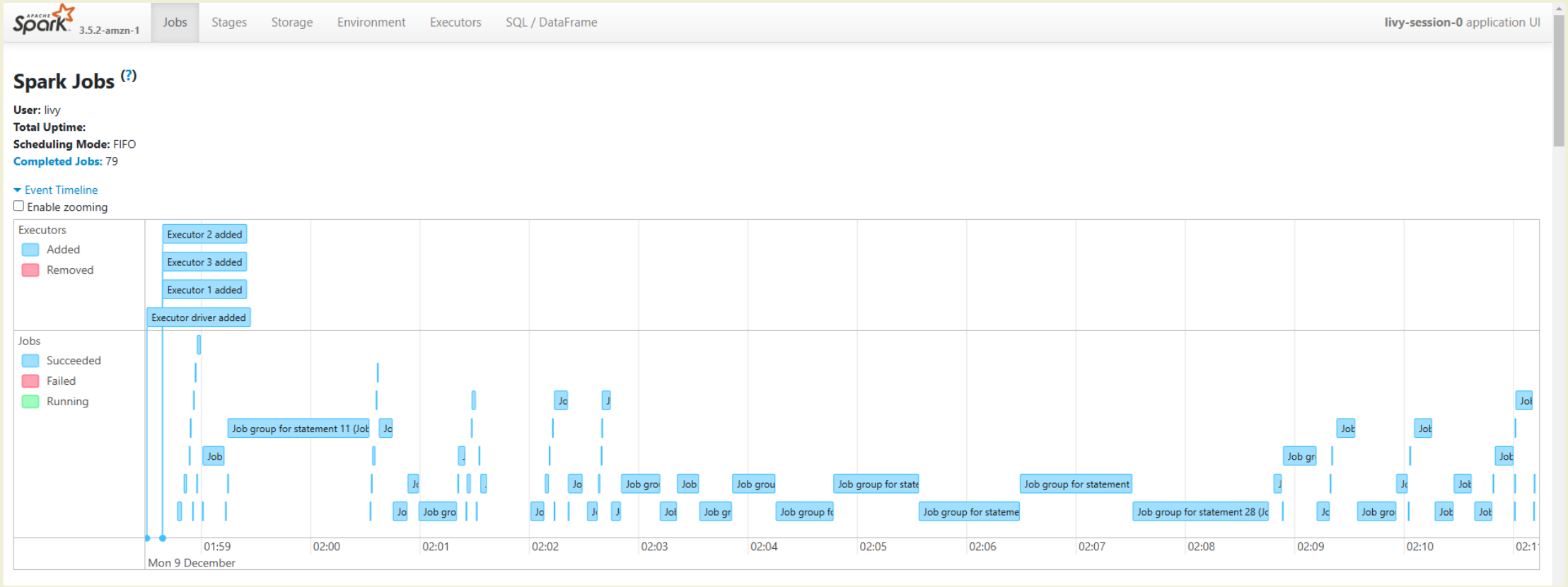
Show 20 entries Search:

Node Labels	Rack	Node State	Node Address	Node HTTP Address	Last health-update	Health-report	Containers	Allocation Tags	Mem Used	Mem Avail	Phys Mem Used %	VCores Used	VCores Avail	Phys VCores Used %	Version
	/default-rack	RUNNING	ip-172-31-17-78.ec2.internal:8041	ip-172-31-17-78.ec2.internal:8042	Mon Dec 09 02:24:41 +0000 2024		1		1.38 GB	10.63 GB	34	1	3	0	3.4.0-amzn-1
	/default-rack	RUNNING	ip-172-31-18-27.ec2.internal:8041	ip-172-31-18-27.ec2.internal:8042	Mon Dec 09 02:24:29 +0000 2024		1		11 GB	1 GB	53	1	3	0	3.4.0-amzn-1
	/default-rack	RUNNING	ip-172-31-18-55.ec2.internal:8041	ip-172-31-18-55.ec2.internal:8042	Mon Dec 09 02:24:30 +0000 2024		1		11 GB	1 GB	51	1	3	0	3.4.0-amzn-1
	/default-rack	RUNNING	ip-172-31-19-132.ec2.internal:8041	ip-172-31-19-132.ec2.internal:8042	Mon Dec 09 02:24:41 +0000 2024		1		11 GB	1 GB	50	1	3	1	3.4.0-amzn-1

Showing 1 to 4 of 4 entries First Previous 1 Next Last

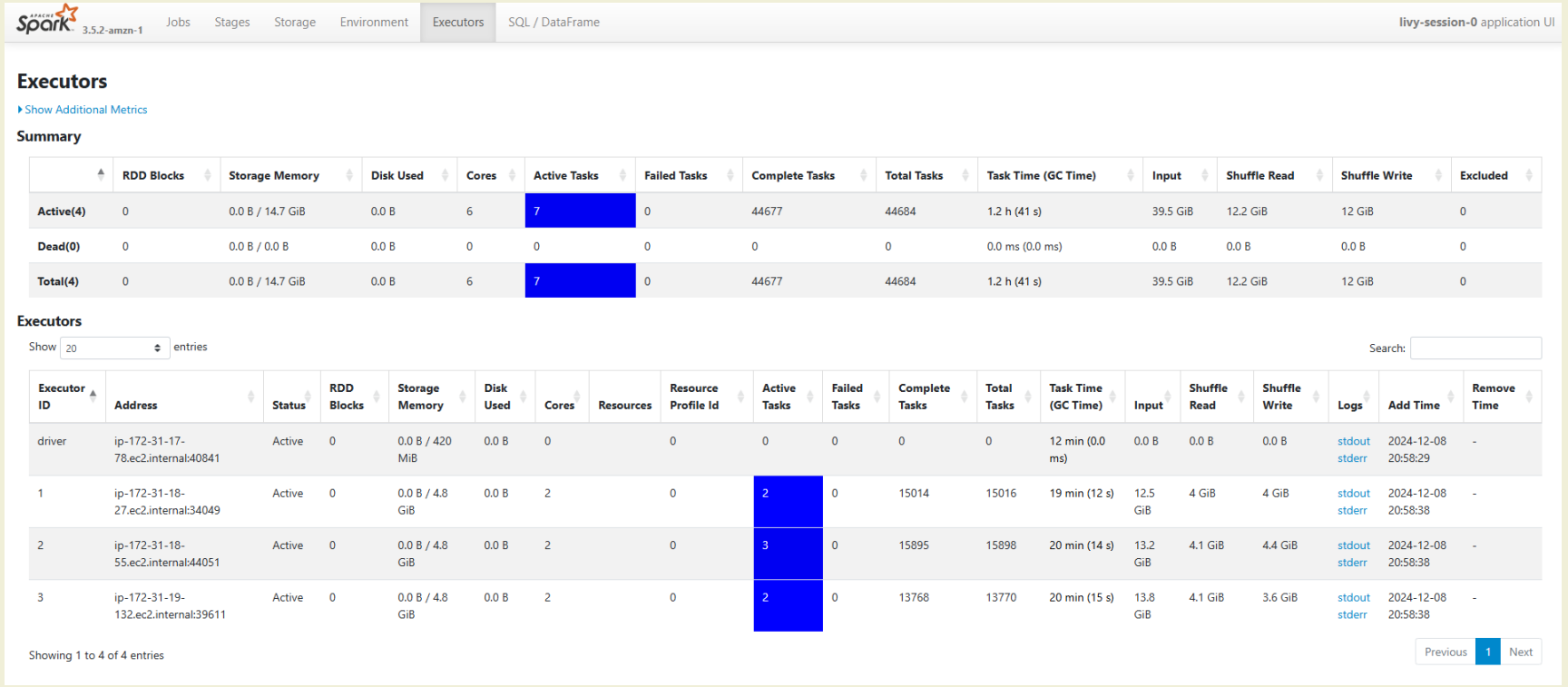
- Memory and vCores were efficiently allocated:
  - Memory Used: ~34.38 GiB (~72% utilization).
  - Physical Cores Used: 4 cores (~100% core efficiency on running tasks).
- Cluster remained stable with no unhealthy nodes or failed tasks during execution.

SPARK JOBS AND PERFORMANCE METRICS



- Total Jobs Executed:
  - 79 (all completed successfully).
  - Execution Time: Majority of jobs completed in milliseconds to seconds.
  - Scheduling: FIFO mode ensured smooth execution of tasks in order of submission

EXECUTORS AND PARALLEL PROCESSING



Spark 3.5.2-amzn-1 | Jobs | Stages | Storage | Environment | Executors | SQL / DataFrame | livy-session-0 application UI

Executors

Show Additional Metrics

Summary

	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write	Excluded
Active(4)	0	0.0 B / 14.7 GiB	0.0 B	6	7	0	44677	44684	1.2 h (41 s)	39.5 GiB	12.2 GiB	12 GiB	0
Dead(0)	0	0.0 B / 0.0 B	0.0 B	0	0	0	0	0	0.0 ms (0.0 ms)	0.0 B	0.0 B	0.0 B	0
Total(4)	0	0.0 B / 14.7 GiB	0.0 B	6	7	0	44677	44684	1.2 h (41 s)	39.5 GiB	12.2 GiB	12 GiB	0

Executors

Show 20 entries Search:

Executor ID	Address	Status	RDD Blocks	Storage Memory	Disk Used	Cores	Resources	Resource Profile Id	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write	Logs	Add Time	Remove Time
driver	ip-172-31-17-78.ec2.internal:40841	Active	0	0.0 B / 420 MiB	0.0 B	0		0	0	0	0	0	12 min (0.0 ms)	0.0 B	0.0 B	0.0 B	stdout stderr	2024-12-08 20:58:29	-
1	ip-172-31-18-27.ec2.internal:34049	Active	0	0.0 B / 4.8 GiB	0.0 B	2		0	2	0	15014	15016	19 min (12 s)	12.5 GiB	4 GiB	4 GiB	stdout stderr	2024-12-08 20:58:38	-
2	ip-172-31-18-55.ec2.internal:44051	Active	0	0.0 B / 4.8 GiB	0.0 B	2		0	3	0	15895	15898	20 min (14 s)	13.2 GiB	4.1 GiB	4.4 GiB	stdout stderr	2024-12-08 20:58:38	-
3	ip-172-31-19-132.ec2.internal:39611	Active	0	0.0 B / 4.8 GiB	0.0 B	2		0	2	0	13768	13770	20 min (15 s)	13.8 GiB	4.1 GiB	3.6 GiB	stdout stderr	2024-12-08 20:58:38	-

Showing 1 to 4 of 4 entries Previous 1 Next

- Executors Overview
  - Active Executors: 4 (including the driver).
  - Task Distribution: Driver executed 420 MiB storage tasks.
  - Task nodes processed ~4.8 GiB each efficiently.

- Execution Time per Task: ~20 minutes, showcasing balanced distribution across nodes.

CONCLUSION

KEY ACHIEVEMENTS

- Successfully processed 3GB transactional data using AWS EMR and PySpark.
- Addressed class imbalance using a combination of majority class reduction and synthetic data generation.
- Engineered key features like FanOut, FanIn, and AvgAmountSent for better laundering pattern detection.
- Achieved robust model performance:
  - ROC: 0.95, F1 Score: 0.90.

FUTURE WORK

- ENHANCEMENTS TO THE MODEL:
  - Cross-Validation: Incorporate cross-validation during training to ensure more robust hyperparameter tuning and model evaluation.
  - Advanced Models: Experiment with advanced algorithms like XGBoost, GNNs to reduce false positives and false negatives for more precise classification.
- CLOUD OPTIMIZATION:
  - Optimize the current cluster configuration for better cost-performance balance.
  - Explore the feasibility of serverless architectures like AWS Lambda for specific workflows.

THANK YOU