CloudWalk

Business Monitoring Intelligence Intern / Daniel Campos França / Universidade Estácio de Sá / 23.07.2024

TABLE OF CONTENTS

SUMMARY

1. INTRODUCTION

This technical report describes how the candidate for the position of monitoring analyst intern at CloudWalk, developed solutions for the two proposed challenges, one being the analysis of anomalies in hypothetical checkout data and the other, the development of a solution for a real related problem. with incident alerts in the company's daily transactions.

Description of Challenges

O **The first challenge** involves analyzing a set of hypothetical checkout data provided in a CSV file. The candidate must develop an analysis model to identify anomalous behavior in operations, present the results obtained and their conclusions.

O **The second challenge** involves developing a transaction incident alert solution using provided CSV files. The candidate must develop a system that monitors transactions in real time, detecting and alerting teams about anomalies in failed, rolled back or denied transactions.

Goals

- Data Analysis: Perform a detailed analysis of the data provided, identifying patterns and anomalies.
- System Development: Implement a monitoring system that detects anomalies in real time and notifies relevant teams.
- Presentation of Results: Communicate clearly and concisely the results of the analysis and the functioning of the developed system, through a structured and detailed presentation.

2. CHALLENGE 1 – Get your hand dirty

The first challenge consists of carrying out an analysis of hypothetical checkout data, using a CSV (Comma-Separated Values) file provided, with the aim of identifying anomalous behaviors in a POS (Point of Sale), presenting the results obtained and the conclusions. This task required skills in interpreting data, creating algorithms to calculate and recognize anomalous data, writing SQL queries, and creating graphical representations to communicate the insights found.

Mathematical methods used

Weighted Mean: The weighted average was calculated by combining the average of the last 7 days and the last month: weighted_mean = $\frac{\left(avg_{last_{week}} \times 7\right)+\left(avg_{las\ month} \times 30\right)}{(7+30)}$

1. Mean: The arithmetic mean of daily values and weekly and monthly averages: mean = $\frac{today+yesterday+same_{day_{last_{week}}}+avg_{last_{week}}+avg_{last_{month}}}{5}$
2. Variance: Variance measures the dispersion of data in relation to the mean. It was calculated using: variance = $\frac{\sum(x-mean)2}{n-1}$
3. Standard Deviation: The standard deviation is the square root of the variance and quantifies the dispersion of the data: std_dev = $\sqrt{variance}$
4. Upper and Lower limits: Determined based on weighted average and standard deviation to identify anomalies.
    - **Upper Limit**: upper_limit = $weighted\_mean\ + (2\ \times\ std\_dev)$
    - **Lower Limit**: lower_limit = $weighte\_mean - (2\ \times\ std\_dev)$
5. Detection of Anomalies within the upper and lower limits: Identification of values that are below 50% of the weighted average as possible anomalies.

Methods used

The **first method** of analysis uses the calculation of standard deviation and upper and lower limits, where the value of K is a constant defined as 2 (K = 2). It should be noted that the value of K determines the width of the control interval, where K could eventually assume the following values: K=1 covers 68% of the data, K=2 covers 95% of the data and K=3 covers 99.7 % of data. This method allows you to identify values that are outside the calculated limits, classifying them as anomalies. Amounts above the limit may be considered fraudulent, as in the case of a stolen card being used for multiple contactless payments at the same point of sale (POS).

The **second method** complements the first method, which is limited to identifying values above or below the established limits. However, there are also values within these limits that can be considered anomalous due to significant deviations from the mean. This may indicate problems with the payment system or signal failures when making transactions. To identify these anomalies, the weighted average was used as a reference value, comparing it with the values of "Today", "Yesterday" and "Same Day Last Week". A value is considered anomalous if it is 50% less than the weighted average.

Checkout 1 analysis

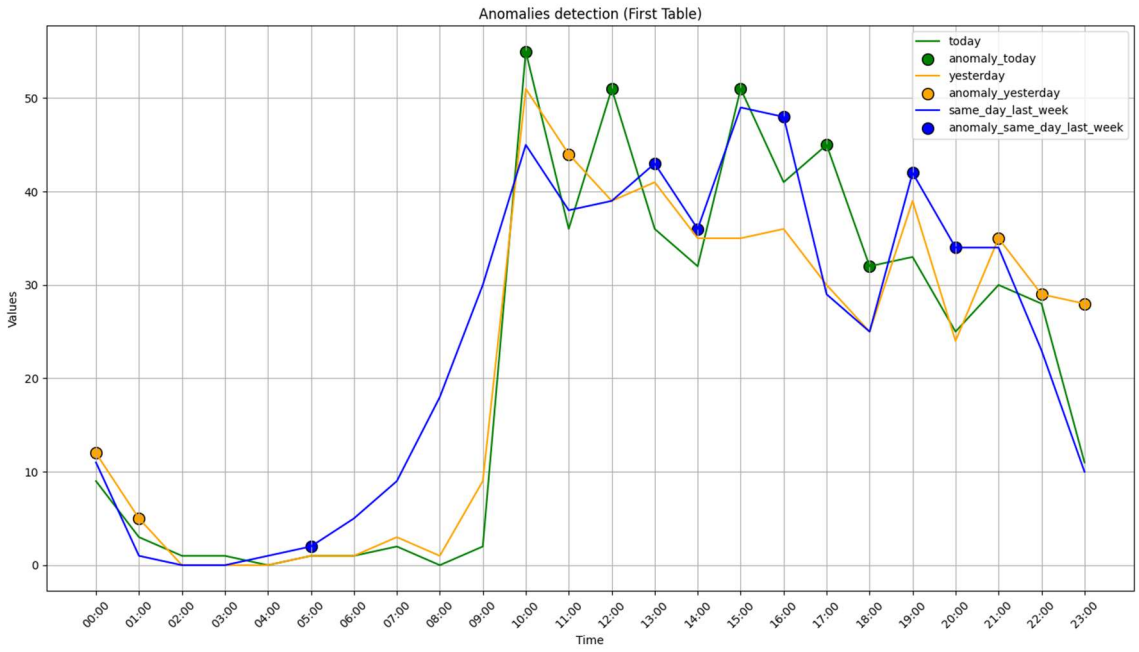
When processing the data as explained previously, the anomalies described by the graphs and tables of Methods 1 and 2 were identified.

The first graph and table represent the anomalies found using Method 1, where 17 anomalies were detected. In the table, cells colored green correspond to anomalies in the "Today" column, yellow cells correspond to anomalies in the "Yesterday" column, and blue cells correspond to anomalies in the "Same Day Last Week" column. The "upper_limit" column indicates the upper limit to consider a value as anomalous and is painted in the color corresponding to the day to facilitate visualization. These values found are represented by their respective colors in the graph.

In the second graph and table, 13 anomalies were identified. The cells have the same colors as the first table: green for "Today", yellow for "Yesterday", and blue for "Same Day Last Week". The columns "today_vs_weighted_mean", "yesterday_vs_weighted_mean" and "same_day_last_week_vs_weighted_mean" are Boolean columns, that is, they contain only true or false values, with the true value indicating an anomaly found.

Analysis method 1

Graphic

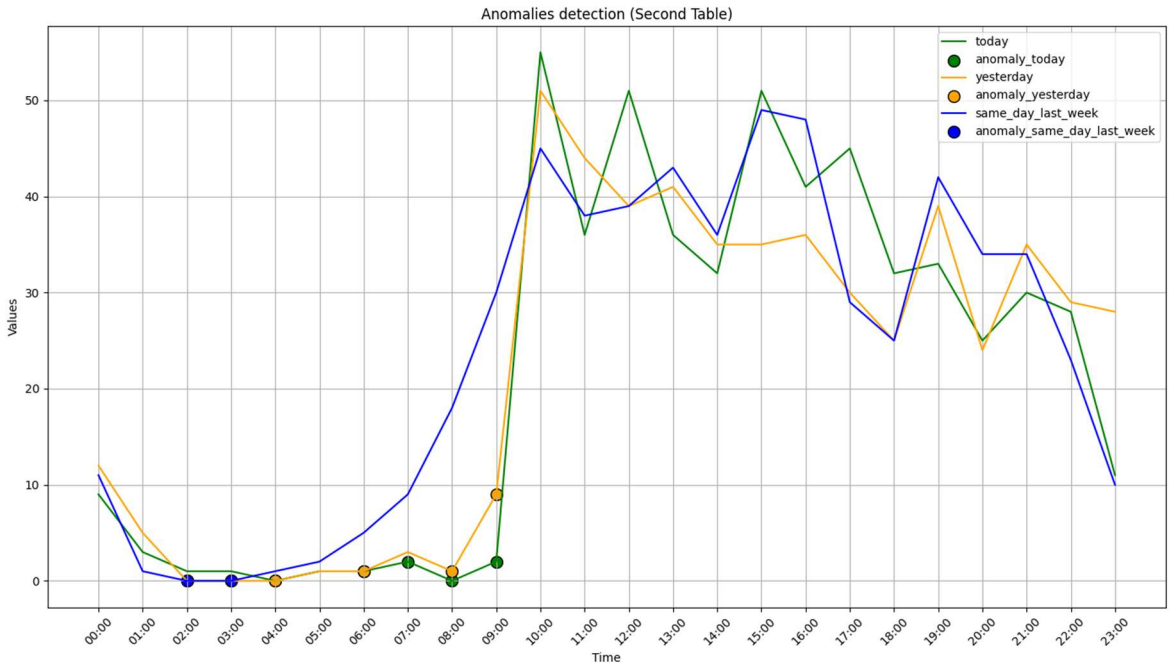


Table

time	today	yesterday	same_day_last_week	avg_last_week	avg_last_month	weighted_mean	mean	variance	std_dev	upper_limit	lower_limit
0:00	9	12	11	6.42	4.85	5.15	8.65	9.07	3.01	11.17	-0.88
1:00	3	5	1	1.85	1.92	1.91	2.55	2.37	1.54	4.99	-1.17
2:00	1	0	0	0.28	0.82	0.72	0.42	0.22	0.47	1.65	-0.21
3:00	1	0	0	0.42	0.46	0.45	0.38	0.17	0.41	1.28	-0.37
4:00	0	0	1	0.42	0.21	0.25	0.33	0.17	0.42	1.08	-0.58
5:00	1	1	2	1.28	0.75	0.85	1.21	0.23	0.48	1.81	-0.11
6:00	1	1	5	2.85	2.28	2.39	2.43	2.72	1.65	5.69	-0.91
7:00	2	3	9	5.57	5.21	5.28	4.96	7.34	2.71	10.7	-0.14
8:00	0	1	18	8.71	10.42	10.1	7.63	54.67	7.39	24.88	-4.69
9:00	2	9	30	20	19.07	19.25	16.01	116.61	10.8	40.84	-2.35
10:00	55	51	45	29.42	28.35	28.55	41.75	150.82	12.28	53.11	3.99
11:00	36	44	38	33.71	28.5	29.49	36.04	32.37	5.69	40.86	18.11
12:00	51	39	39	27.57	25.42	25.83	36.4	106.3	10.31	46.45	5.21
13:00	36	41	43	25.85	24.21	24.52	34.01	74.07	8.61	41.73	7.31
14:00	32	35	36	26.14	25.21	25.39	30.87	24.76	4.98	35.34	15.43
15:00	51	35	49	28.14	27.71	27.79	38.17	125.49	11.2	50.2	5.39
16:00	41	36	48	27.71	25.64	26.03	35.67	86.13	9.28	44.59	7.47
17:00	45	30	29	20.42	22.28	21.93	29.34	93.8	9.68	41.3	2.56
18:00	32	25	25	21.57	18.28	18.9	24.37	25.98	5.1	29.1	8.71
19:00	33	39	42	22.14	18.67	19.33	30.96	104.88	10.24	39.81	-1.16
20:00	25	24	34	17.42	18.92	18.64	23.87	42.5	6.52	31.68	5.6
21:00	30	35	34	18.71	17.57	17.79	27.06	69.91	8.36	34.51	1.06
22:00	28	29	23	15.42	15.64	15.6	22.21	42.38	6.51	28.62	2.58
23:00	11	28	10	9.57	8.75	8.91	13.46	66.69	8.17	25.24	-7.43

Analysis method 2

Graphic



Table

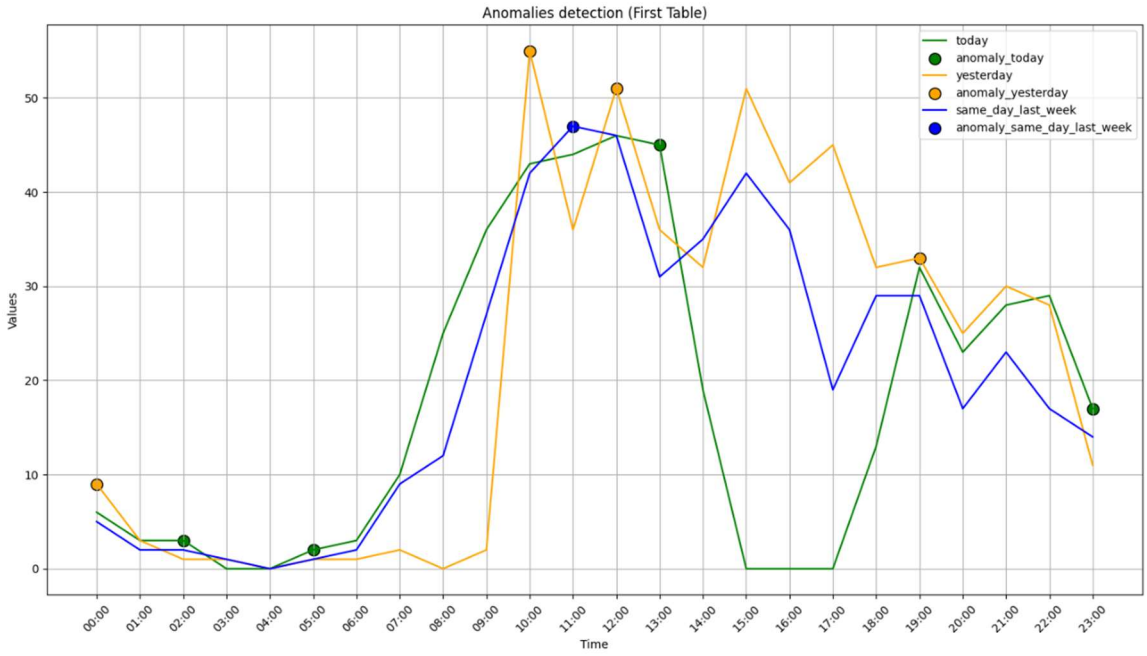
time	today	yesterday	same_day_last_week	weighted_mean	today_vs_weighted_meany	yesterday_vs_weighted_mean	same_day_last_week_vs_weighted_mean
0:00	9	12	11	5.15	FALSE	FALSE	FALSE
1:00	3	5	1	1.91	FALSE	FALSE	FALSE
2:00	1	0	0	0.72	FALSE	TRUE	TRUE
3:00	1	0	0	0.45	FALSE	TRUE	TRUE
4:00	0	0	1	0.25	TRUE	TRUE	FALSE
5:00	1	1	2	0.85	FALSE	FALSE	FALSE
6:00	1	1	5	2.39	TRUE	TRUE	FALSE
7:00	2	3	9	5.28	TRUE	FALSE	FALSE
8:00	0	1	18	10.1	TRUE	TRUE	FALSE
9:00	2	9	30	19.25	TRUE	TRUE	FALSE
10:00	55	51	45	28.55	FALSE	FALSE	FALSE
11:00	36	44	38	29.49	FALSE	FALSE	FALSE
12:00	51	39	39	25.83	FALSE	FALSE	FALSE
13:00	36	41	43	24.52	FALSE	FALSE	FALSE
14:00	32	35	36	25.39	FALSE	FALSE	FALSE
15:00	51	35	49	27.79	FALSE	FALSE	FALSE
16:00	41	36	48	26.03	FALSE	FALSE	FALSE
17:00	45	30	29	21.93	FALSE	FALSE	FALSE
18:00	32	25	25	18.9	FALSE	FALSE	FALSE
19:00	33	39	42	19.33	FALSE	FALSE	FALSE
20:00	25	24	34	18.64	FALSE	FALSE	FALSE
21:00	30	35	34	17.79	FALSE	FALSE	FALSE
22:00	28	29	23	15.6	FALSE	FALSE	FALSE
23:00	11	28	10	8.91	FALSE	FALSE	FALSE

Checkout 2 analysis

When processing data from checkout 2 in the same way as data from checkout 1, it can be seen that in the first process, 9 anomalies were identified and in the second process, 10 anomalies were identified.

Analysis method 1

Graphic

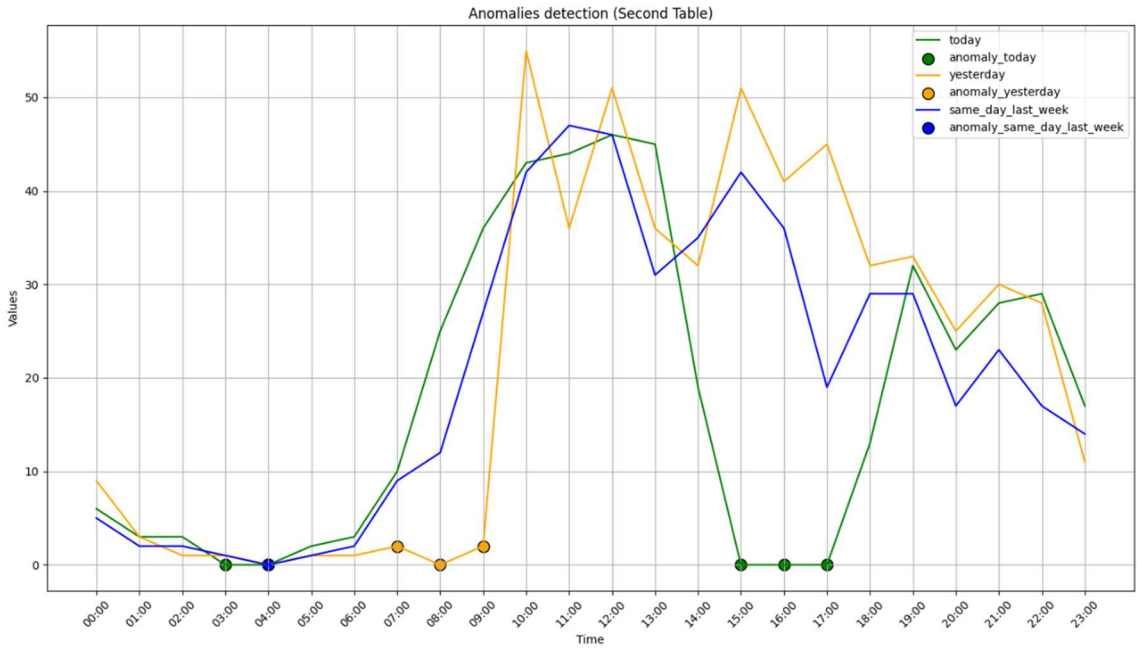


Table

time	today	yesterday	same_day_last_week	avg_last_week	avg_last_month	weighted_mean	mean	variance	std_dev	upper_limit	lower_limit
0:00	6	9	5	5	4.92	4.94	5.98	3.04	1.74	8.42	1.45
1:00	3	3	2	2	1.92	1.94	2.38	0.32	0.56	3.06	0.81
2:00	3	1	2	0.42	0.75	0.69	1.43	1.11	1.06	2.8	-1.42
3:00	0	1	1	0.42	0.46	0.45	0.58	0.18	0.43	1.31	-0.4
4:00	0	0	0	0.14	0.21	0.2	0.07	0.01	0.1	0.39	0
5:00	2	1	1	0.71	0.71	0.71	1.08	0.28	0.53	1.77	-0.35
6:00	3	1	2	1.42	2.1	1.97	1.9	0.58	0.76	3.49	0.45
7:00	10	2	9	3	5.03	4.65	5.81	12.69	3.56	11.77	-2.48
8:00	25	0	12	3.71	9.82	8.66	10.11	92.14	9.6	27.86	-10.53
9:00	36	2	27	10.14	17.64	16.22	18.56	180.34	13.43	43.08	-10.64
10:00	43	55	42	26.14	28.57	28.11	38.94	138.79	11.78	51.67	4.55
11:00	44	36	47	25	28.28	27.66	36.06	91.4	9.56	46.78	8.54
12:00	46	51	46	24	25.89	25.53	38.58	159.5	12.63	50.79	0.27
13:00	45	36	31	20.28	24.17	23.43	31.29	95.54	9.77	42.98	3.89
14:00	19	32	35	19.57	24.89	23.88	26.09	52.13	7.22	38.32	9.44
15:00	0	51	42	22.427	27.78	26.77	28.64	384.51	19.61	65.99	-12.45
16:00	0	41	36	21.57	25.53	24.78	24.82	253.47	15.92	56.62	-7.06
17:00	0	45	19	17.71	22.67	21.73	20.88	258.63	16.08	53.9	-10.43
18:00	13	32	29	16.85	18.46	18.16	21.86	67.24	8.2	34.56	1.76
19:00	32	33	29	18	18.21	18.17	26.04	54.67	7.39	32.96	3.38
20:00	23	25	17	12.14	18.53	17.32	19.13	25.8	5.08	27.48	7.16
21:00	28	30	23	14.85	17.82	17.26	22.73	41.73	6.46	30.18	4.34
22:00	29	28	17	12.71	15.5	14.97	20.44	56.6	7.52	30.02	-0.08
23:00	17	11	14	8.28	8.75	8.66	11.81	13.55	3.68	16.02	1.3

Analysis method 2

Graphic



Table

time	today	yesterday	same_day_last_week	weighted_mean	today_vs_weighted_meany	yesterday_vs_weighted_mean	same_day_last_week_vs_weighted_mean
0:00	6	9	5	4.94	FALSE	FALSE	FALSE
1:00	3	3	2	1.94	FALSE	FALSE	FALSE
2:00	3	1	2	0.69	FALSE	FALSE	FALSE
3:00	0	1	1	0.45	TRUE	FALSE	FALSE
4:00	0	0	0	0.2	TRUE	TRUE	TRUE
5:00	2	1	1	0.71	FALSE	FALSE	FALSE
6:00	3	1	2	1.97	FALSE	FALSE	FALSE
7:00	10	2	9	4.65	FALSE	TRUE	FALSE
8:00	25	0	12	8.66	FALSE	TRUE	FALSE
9:00	36	2	27	16.22	FALSE	TRUE	FALSE
10:00	43	55	42	28.11	FALSE	FALSE	FALSE
11:00	44	36	47	27.66	FALSE	FALSE	FALSE
12:00	46	51	46	25.53	FALSE	FALSE	FALSE
13:00	45	36	31	23.43	FALSE	FALSE	FALSE
14:00	19	32	35	23.88	FALSE	FALSE	FALSE
15:00	0	51	42	26.77	TRUE	FALSE	FALSE
16:00	0	41	36	24.78	TRUE	FALSE	FALSE
17:00	0	45	19	21.73	TRUE	FALSE	FALSE
18:00	13	32	29	18.16	FALSE	FALSE	FALSE
19:00	32	33	29	18.17	FALSE	FALSE	FALSE
20:00	23	25	17	17.32	FALSE	FALSE	FALSE
21:00	28	30	23	17.26	FALSE	FALSE	FALSE
22:00	29	28	17	14.97	FALSE	FALSE	FALSE
23:00	17	11	14	8.66	FALSE	FALSE	FALSE

Conclusion – Challenge 1 Get your hands dirty

Data analysis, as detailed in the methods presented, revealed significant information about the anomalies in the analyzed data, as reported below:

1. Analysis of Identified Anomalies:
    - Values Above Upper Limits: Values that exceed upper limits may suggest significant local events, such as an abrupt increase in the number of transactions. It must be assessed whether such peaks are associated with special events, promotions or periods of high demand.
    - Anomalous Values that tend to Zero: Anomalous values equal to zero may indicate possible operational problems, such as a lack of energy at the location, which would result in the temporary closure of the establishment, or unforeseen events that are not related to the payment system.
2. Analysis of checkout tables 1 and 2 provided additional insights that may be useful for the investigation:
    - 24-Hour Transactions: The location in question records transactions over 24 hours, suggesting that it is a 24-hour convenience store. The largest number of customers frequent the establishment between approximately 8 am and 10 pm.
    - Investigation of Fraudulent Transactions: To check the possibility of fraudulent transactions, it is recommended to examine the location where the POS (point of sale) is installed. Some locations may have higher fraud rates than others, and analyzing transaction history can help identify suspicious patterns.

Finally, the initial analysis provides a solid basis for understanding the detected anomalies. Subsequent investigations should focus on correlating these anomalies with specific events and evaluating possible operational or safety causes.

3. Challenge 2 – Solve the problem

The challenge presented involves the implementation of a real-time alert monitoring system for financial transactions, using data from a CSV file containing information about approved, denied, failed and reversed transactions, with the aim of identifying anomalous behaviors and generating automatic notifications for the responsible teams.

This task required data interpretation skills, creating algorithms to calculate and recognize anomalies, writing SQL queries and developing graphical representations to communicate the insights found. Additionally, it was necessary to implement an endpoint to receive transaction data and return alert recommendations, ensuring that the system could operate efficiently and effectively.

Mathematical methods used

1. Mean
2. Standard Deviation
3. Z-score: is a standardized way of measuring how much a value deviates from the mean of a set of data: $z = \frac{x-\mu}{\sigma}$
    - x: The individual value of the dataset being analyzed.
    - μ: The average of the data set values.
    - σ: The standard deviation of the dataset values.

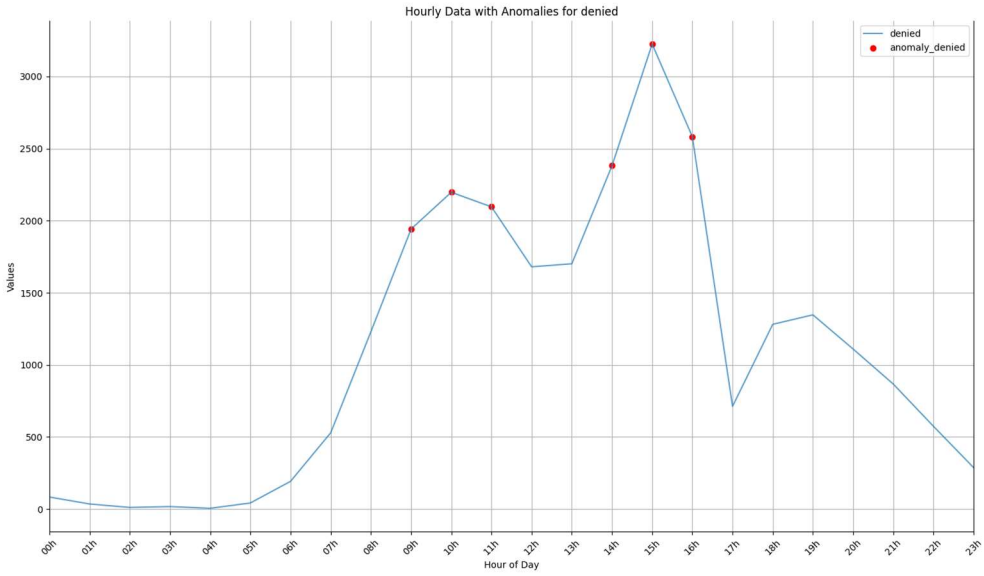
Methods used

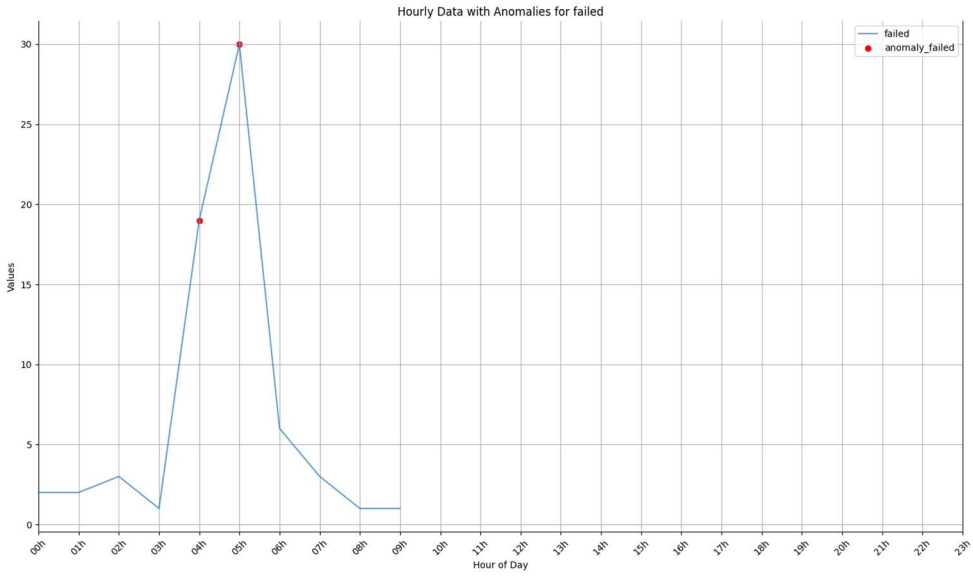
The method used to identify anomalies in the "transactions" file consisted of adding up all values by status in each of the 24 hours presented. After this sum, the simple average of the values for the "denied", "reversed" and "failed" statuses was calculated, resulting in 24 metrics for each status. A z-score of 0.7 was defined, as, after analyzing the processed data, it was found that values above this threshold are not recognized as anomalies, as they are noticeable when inspecting the table.

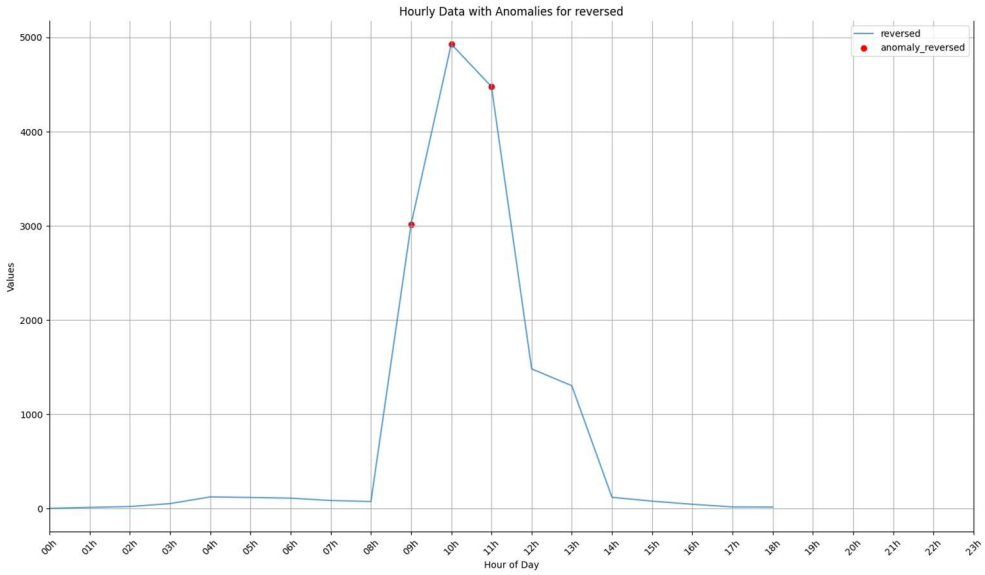
Transaction 1 analysis

When processing the data from the "transactions 1" file, 6 anomalous values were identified with the status "denied", 2 with the status "failed" and 3 with the status "reversed". In the table, anomalous values were highlighted with specific colors to facilitate visualization: yellow for "denied", green for "failed" and blue for "reversed". Furthermore, it was found that all anomalous values had a z-score greater than 0.7.

Graphics






Hourly Data with Anomalies for reversed
reversed
anomaly_reversed
Values
Hour of Day
5000
4000
3000
2000
1000
0
00h
01h
02h
03h
04h
05h
06h
07h
08h
09h
10h
11h
12h
13h
14h
15h
16h
17h
18h
19h
20h
21h
22h
23h

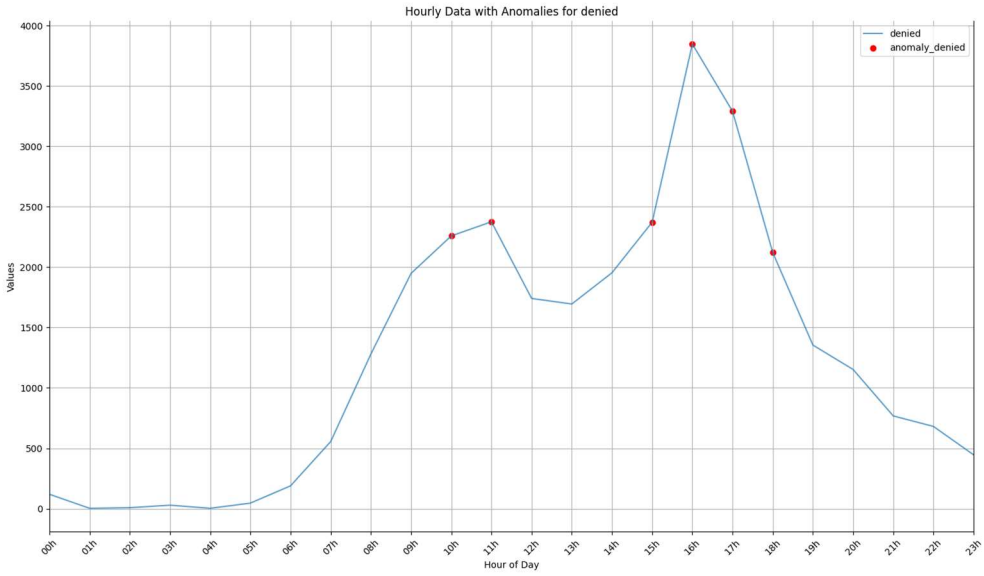
Table

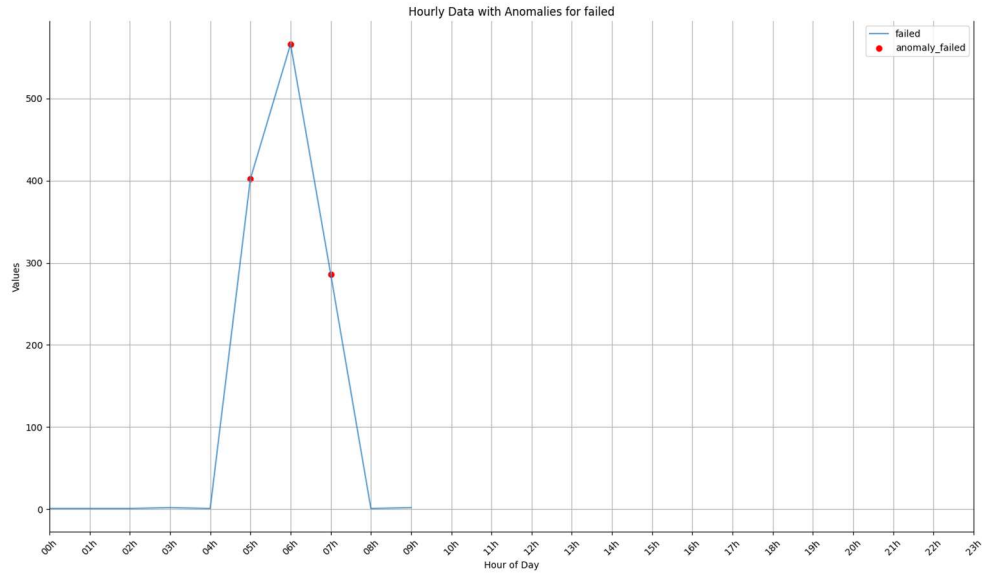
		Z-SCORE		Z-SCORE		Z-SCORE
	denied	denied	failed	failed	reversed	reversed
0h	83.00	-1.05	0.00	-0.40	3.00	-0.47
1h	35.00	-1.10	0.00	-0.40	0.00	-0.47
2h	11.00	-1.12	0.00	-0.40	0.00	-0.47
3h	17.00	-1.12	0.00	-0.40	0.00	-0.47
4h	5.00	-1.13	0.00	-0.40	0.00	-0.47
5h	42.00	-1.09	0.00	-0.40	0.00	-0.47
6h	192.00	-0.93	0.00	-0.40	13.00	-0.46
7h	529.00	-0.58	0.00	-0.40	22.00	-0.46
8h	1230.00	0.15	2.00	-0.12	53.00	-0.43
9h	1942.00	0.89	0.00	-0.40	124.00	-0.38
10h	2197.00	1.15	2.00	-0.12	118.00	-0.39
11h	2096.00	1.05	0.00	-0.40	111.00	-0.39
12h	1680.00	0.62	3.00	0.02	86.00	-0.41
13h	1701.00	0.64	1.00	-0.26	75.00	-0.42
14h	2381.00	1.35	19.00	2.30	3017.00	1.65
15h	3225.00	2.23	30.00	3.87	4929.00	2.99
16h	2582.00	1.56	6.00	0.45	4478.00	2.67
17h	712.00	-0.39	0.00	-0.40	1483.00	0.57
18h	1281.00	0.20	3.00	0.02	1304.00	0.44
19h	1347.00	0.27	0.00	-0.40	120.00	-0.39
20h	1110.00	0.02	1.00	-0.26	79.00	-0.42
21h	867.00	-0.23	0.00	-0.40	46.00	-0.44
22h	574.00	-0.54	0.00	-0.40	18.00	-0.46
23h	287.00	-0.83	1.00	-0.26	17.00	-0.46

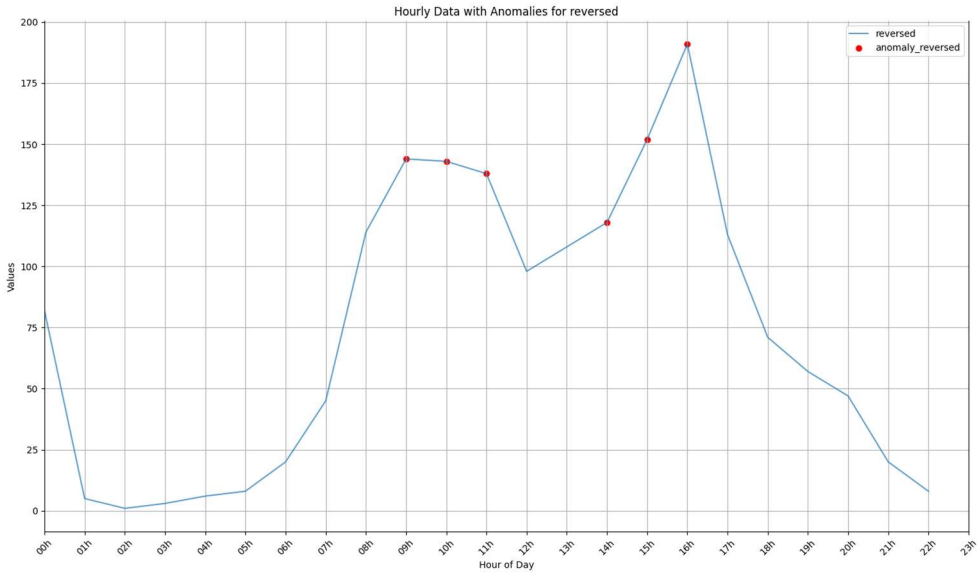
Transaction 2 analysis

When processing the data from the "transactions 1" file, 6 anomalous values were identified with the status "denied", 3 with the status "failed" and 6 with the status "reversed". In the table, anomalous values were highlighted with specific colors to facilitate visualization: yellow for "denied", green for "failed" and blue for "reversed". Furthermore, it was found that all anomalous values had a z-score greater than 0.7.

Graphics






Hourly Data with Anomalies for reversed
reversed
anomaly_reversed
Values
Hour of Day
200
175
150
125
100
75
50
25
0
00h
01h
02h
03h
04h
05h
06h
07h
08h
09h
10h
11h
12h
13h
14h
15h
16h
17h
18h
19h
20h
21h
22h
23h

Table

		Z-SCORE		Z-SCORE		Z-SCORE
	denied	denied	failed	failed	reversed	reversed
0h	119.00	-1.03	0.00	-0.36	82.00	0.19
1h	3.00	-1.13	0.00	-0.36	5.00	-1.10
2h	8.00	-1.13	0.00	-0.36	1.00	-1.17
3h	29.00	-1.11	0.00	-0.36	3.00	-1.13
4h	3.00	-1.13	0.00	-0.36	0.00	-1.18
5h	46.00	-1.09	0.00	-0.36	6.00	-1.08
6h	189.00	-0.96	0.00	-0.36	8.00	-1.05
7h	556.00	-0.63	0.00	-0.36	20.00	-0.85
8h	1281.00	0.02	1.00	-0.35	45.00	-0.43
9h	1949.00	0.62	0.00	-0.36	114.00	0.73
10h	2259.00	0.90	1.00	-0.35	144.00	1.23
11h	2376.00	1.01	1.00	-0.35	143.00	1.22
12h	1741.00	0.43	2.00	-0.34	138.00	1.13
13h	1695.00	0.39	1.00	-0.35	98.00	0.46
14h	1954.00	0.62	0.00	-0.36	108.00	0.63
15h	2373.00	1.00	0.00	-0.36	118.00	0.80
16h	3848.00	2.33	402.00	2.38	152.00	1.37
17h	3291.00	1.83	566.00	3.49	191.00	2.02
18h	2122.00	0.78	286.00	1.59	113.00	0.71
19h	1355.00	0.09	1.00	-0.35	71.00	0.01
20h	1153.00	-0.10	0.00	-0.36	57.00	-0.23
21h	768.00	-0.44	0.00	-0.36	47.00	-0.39
22h	681.00	-0.52	2.00	-0.34	20.00	-0.85
23h	446.00	-0.73	0.00	-0.36	8.00	-1.05

Alert system

An alert system was developed to notify the responsible team about possible anomalies using the CSV files provided. This system sends an email whenever a file is processed, which can contain up to three images, each corresponding to a specific status (denied, failed and reversed). The email also includes a detailed list, organized by status, that reports times identified as anomalous.

Conclusion – Challenge 2 Solve the problem

The analyzed data highlights the times when anomalies were identified in both files. However, it is essential that the team that received the alert carry out further investigation, examining all transactions at these times to identify the source of the problem.

Additionally, a “/metrics” endpoint was developed for integration with Prometheus and Grafana, with the aim of allowing real-time monitoring and updating of graphs. However, problems arose related to the parameterization of exported data and the configuration of the dashboard in Grafana, which prevented the full implementation of these solutions.

4. FINAL CONCLUSION

This technical report presents the analyses, developments and conclusions inherent to each of the challenges proposed by CloudWalk for the position of monitoring analyst intern, which involved the analysis of anomalies in hypothetical checkout data and the development of a transaction monitoring system financial services for anomaly detection.

To this end, data analysis and systems development techniques were used, where it was possible to identify suspicious patterns and generate real-time alerts for the responsible team, providing greater security in financial operations.

In an initial analysis, it is observed that if the developed system were put into production, it would result in the analysis of millions of daily transactions, resulting in a high processing cost. A potential solution is to implement a prior risk classification system for each POS, which would be scalable and based on factors such as customer profile analysis, location of the enterprise (state, municipality and neighborhood), type of business and transaction history.

For example, a jewelry store in a high-end shopping mall in São Paulo would have a different risk rating than a supermarket in a peripheral neighborhood in Rio de Janeiro. The jewelry store, with few transactions and high security controls, would have a lower risk of fraud, while the supermarket, with many transactions and a greater number of employees, would have a higher risk.

Based on this classification, low-risk POS could have a less thorough analysis, focused only on discrepant transactions, while high-risk POS would require a more

detailed investigation of all transactions. This approach would allow for more efficient and scalable analysis, optimizing technical and financial resources and maintaining the security of financial transactions.