Data Quality Assurance in Credit Card Fraud Prevention

Yuqing Yao

yqelisa@gmail.com

Abstract

In 2023, credit card fraud cases soared to 425,977, up 53% from 2019 driven by the surge in online shopping amid the COVID-19 pandemic. This underscores the pressing need for advancing Fraud Prevention techniques, heavily relying on data analysis and modeling.

While modeling techniques for fraud detection are widely discussed in academia, data quality assurance for it is a less popular topic. However, with the challenge of data imbalance and the manual components in fraud labeling, data quality assurance becomes even more important to ensure the quality of fraud detection. This article discusses types, causes, and solutions for data quality assurance in fraud detection, addressing:

1. Impact of data quality on credit card fraud detection
2. Challenges specific to fraud detection data quality assurance
3. Types of datasets used in this field
4. Tools and procedures
5. Causes of data quality issues based on different dimensions
6. Solutions for data quality issues

Besides reviewing recent academic research papers, this article incorporates industry research reports, technical blogs from leading FinTech companies, and vendor product specifications. It focuses on actionable insights for data quality assurance in credit card fraud detection.

Biography

Yuqing Yao currently works as a Senior Fraud Analyst at a FinTech company, Klarna. She uses statistical approaches and visualization to detect fraud patterns and program them in the fraud decisioning system. Prior to that, she worked as Assistant Director in Data Operations at Moody's Analytics, leading the data processing and reporting of its credit data consortium. She started her career in the Treasury team at Stripe, doing liquidity forecasting and monitoring. Her career interests have been in data science, analytics, and risk management. In her spare time, she actively volunteers for organizing AI/ML meetups in San Francisco. Besides, she is also a long-term Toastmasters.

1 Introduction

1.1 Why should we care about Credit Card Fraud Prevention

Credit cards brought us convenience and enabled us to use leverage, thus becoming an essential part of the economy. As of 2023, credit card is the top payment method by volume and the second largest payment method for online transactions by volume (Worldpay 2024).

Fraud prevention protects customers from bad actors. Based on the U.S. Federal Trade Commission (FTC) report, consumers incurred over $10 billion fraud losses in 2023 (Federal Trade Commission 2024), which is approximately 150,000 US households' annual expenditure in that year. We can't stop fraudsters from having ill intents, but we can protect the financial institutions and consumers from those bad actions. Therefore, keeping fraud prevention methods and technologies up to date is how we win this cat-and-mouse game.

1.2 The role of data in Credit Card Fraud Prevention

Fraud prevention heavily relies on data to differentiate fraudsters from genuine customers. Data is used in multiple ways in the lifecycle of fraud prevention:

- **Behavioral analysis:** Use customer activities to find loopholes in product design that let fraudsters in.
- **Detect fraud attacks:** Detect abnormal volumes and trigger alerts.
- **Case studies:** Extract patterns from fraudsters' behaviors to understand how they conduct frauds.
- **Real-time decisioning:** Use data to inform which customer or transaction should be blocked.

1.3 Why is data quality important in Credit Card Fraud Prevention

1.3.1 Data Driven Decisioning

According to a survey in the Insurance Fraud space (SAS 2021), 64% of the respondents think that poor data quality is the most significant challenge for implementing fraud detection technology. Similarly, fraud prevention for credit cards requires a high level of automation, because of the large volume of sign ups and transactions. How we tell the fraudsters or fraudulent transactions apart from the legitimate customers and transactions is largely dependent on recognizing the patterns of these two groups. This means that the quality of the data we collect affects the quality of our decisions.

1.3.2 Class Imbalance

Identifying frauds from non-frauds is a classification problem. Fraudsters are usually a very small percentage of the sign up or transaction attempts making the sample size of this class much smaller than the other class, which creates a big challenge for classification models. Despite having ways to handle the imbalance, upsampling for example, the lack of data still makes training sensitive to small data errors in the fraud class.

1.3.3 Fraud as Anomaly

Anomaly Detection is also widely used in fraud detection. By definition, anomalies usually have extreme values or weir patterns compared to normal observations. If a normal customer's data has errors while being collected, the customer could be wrongly categorized as a fraudster, even though the actual values are normal. This increases False Positives and creates negative customer experience.

1.4 Scope of this Paper

This paper reviews academic papers about data quality and fraud analytics, to find the intersection of these two topics, which is not widely explored but practical. It sorts the problems and solutions for data quality in credit card fraud prevention into a logical framework, making them easy to comprehend. It also synthesizes the latest best practices from technical blogs of leading global FinTech companies and industry conferences to map out different solutions.

The rest of the paper is organized as follows: Section 2 lays out the types of datasets used in Credit Card Fraud Prevention to show the objects of quality assurance. Section 3 lists the tools (technologies and processes) used in this area. Section 4 discusses the commonly seen data quality issues and the cause of them. Section 5 proposes solutions to the different data quality issues. Section 6 summarizes the work and suggests future improvements.

2 Types of Datasets

Fraud prevention uses different types of data to paint a holistic picture of the customer. Each type of data has some unique data quality problems.

2.1 Customers

The first step of blocking credit card fraud is blocking fraudsters from signing up. This involves collecting customer information and verifying if it matches the real customers. Typical fields being collected are:

- First Name and Last Name
- Date of Birth
- National Identity Number
- Billing / Mailing Address
- Occupation
- Email
- Phone Number

Fraudsters could steal a real customer's identity and sign-up credit cards on their behalf, leaving the real customer liable for the credit card debt, or the financial institution to eat the cost. This type of fraud is called Identity Theft.

Alternatively, fraudsters could make up fake identities and try to fool the financial institution to issue credit cards. For example, making up a fake name, but matching with corresponding identity documents to make the sign-up material seem legitimate. This is called Synthetic Identity.

To block these types of fraud, we compare the self-reported information during the sign-up process with the data previously on file, or with external bureaus, to see if they are consistent.

2.2 Transactions

If the customer's account is hacked, or the customer's card information (card number and CVV) is stolen, fraudsters could use a card from a legitimate customer to make purchases. These frauds are called Account Takeover and Card Skimming, correspondingly.

Sometimes a customer does not want to pay back the credit card debt, the customer claims someone else (a fraudster) made those transactions, hoping to shift the loss to the financial institution, this is called First Party Fraud.

In both situations, data of the specific transactions is critical in identifying which transactions are fraudulent and which ones are not. The following data points are interesting to fraud prevention:

- Card Number
- Card Verification Value (CVV)
- Billing Address
- Expiration Date
- Transaction Amount
- Transaction Currency
- Merchant Name
- Transaction Description
- Entry Mode

2.3 Behaviors

Besides looking at the customer information at sign up and individual transaction information, we also look into customer behaviors over time.

For example:

- Log in Device
- Log in Time
- Log in Location
- Purchased Items
- Average Purchase Amount

When we observe a different pattern compared to this customer's previous behavior, we would suspect it's a fraudulent transaction or log-in.

2.4 External Data Sources

There are external data providers that supply data points not directly collected through the financial institution's interactions with the customer. For example:

- Credit bureau data
- Background check data
- Social media data
- Fraudulent reports from other financial institutions

2.5 Datasets Summary

Each aforementioned set of data alone is not enough for identifying fraud. The more comprehensive the datasets are, the fuller picture we have of the sign up or transactions. Used in isolation, each of them has their own set of problems. However, them being used together creates a new set of problems, which will be explained in section 5.2.

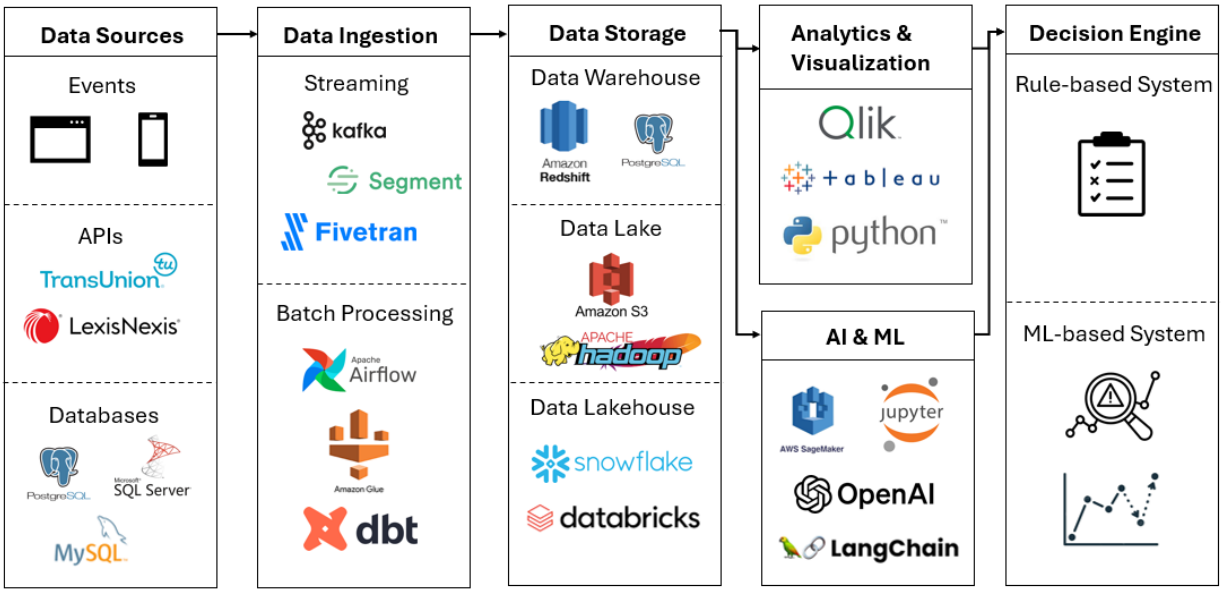
3 Data Stack

A Data Stack is a collection of technologies and processes that enable data to be collected, processed, and then used. Fraud prevention shares a similar data stack of many other e-commerce analytical problems. Understanding the data stack helps us identify where the data quality problems are introduced.

3.1 Data Stack Diagram

The below diagram shows how each of the components interact with each other, and where they are positioned in the workflow for fraud prevention.

- Data enters the system from sources through ingestion tools.
- Then the data is pushed to be stored in some structure, Data Warehouse, Data Lake, or Lakehouse.
- From there, we can consume the data by visualizing or generating predictions.
- Finally, we use the insights from analytics or modeling to trigger actions in the Decision Engine, where sign-up or transaction attempts are marked with “accept” or “reject”.



3.2 Data Sources

The previous section, we looked at the types of data used in Fraud Prevention. Some types are collected directly from interacting with the customer, and some come from external parties:

- **Customers**: customer submission through online forms
- **Transactions**: collected from the card network
- **Behaviors**: customer frontend events
- **Merchants**: collected from the card network
- **External Data Sources**: integration with data vendors and credit bureaus

3.3 Data Ingestion

Data ingestion is the process of obtaining, importing, and processing data from multiple sources into a single storage place. The purpose of data ingestion is to standardize and centralize data for different teams in the organization to use. Data storage locations will be explained in the next section.

Depending on whether data is processed as it is generated, or first being loaded then being processed altogether, we have batch processing or streaming. Common choices among these two are:

- **Streaming**: real-time fraud decisioning data (transactions), front-end events
- **Batch processing**: non-time-sensitive fraud variables (statistics over a certain time window, segment benchmarks), analytics, and reporting

3.4 Data Storage

Based on the needs of the organization and nature of data, there are following forms of storage (Striim 2024):

- **Data Warehouse**: a unified data repository for storing a large amount of data, usually in structured format.
- **Data Lake**: a centralized and flexible data repository that stores both structured and unstructured data in its raw form.
- **Data Lakehouse**: A hybrid of warehouse and lake, which enables storage of both structured, semi-structured, and unstructured data. It's built upon data lakes that stores all the raw datasets, and then applied ACID transactional processes that are data warehouses' features.

Because fraud prevention involves a fair amount of both analytics and Machine Learning, data lakehouse is gaining popularity in major FinTech companies. Thus, credit card fraud detection is also often conducted based on the lakehouse architecture.

3.5 Data Analytics

Analytics include data cleaning, transformation, and calculation. This is when insights are drawn from data, and then guide actions.

Fraud detection involves a lot of deep dives and ad hoc analysis. These types of analysis are usually done in Jupyter Notebooks or spreadsheets. Lakehouses usually support SQL querying, and that is how analytics is done in many cases.

In some cases, vendor data is only accessible through the vendor environment, which makes these vendor software applications a part of the analytics component of the data stack.

3.6 Machine Learning and Artificial Intelligence

More advanced fraud prevention systems use Machine Learning to support decision-making. This usually includes functions for:

- Feature Engineering
- Model Training and Evaluation
- Model Deployment
- Continuous Monitoring and Updating

3.7 Visualization and Reporting

Visualization serves multiple purposes for fraud prevention:

- Build intuition on fraud trends and patterns
- Detect anomalies and surface potential fraud attacks
- Showcase fraud prevention engine and algorithm performance to stakeholders

The tools for visualization include:

- Dashboards
- Jupyter Notebooks
- Reports (usually in PDF format)

3.8 Decision Engine

This is where insights are turned into actions. It's the system that streams transactions or card opening applications data, go through a set of rules defined by the card issuer to calculate the fraud risk, and come up with either accept or reject decision based on that.

The following categories of decision engines are common in credit card fraud prevention:

- **Rule-based systems**: Analysts discover fraud patterns using analytics, and then define a set of rules each transaction or sign-up needs to go through. This is usually easy to comprehend, because the thresholds and conditions used in these rules have context, for example, a fraud attack, or a data breach.
- **Anomaly detection**: It's also relatively easy to comprehend, because it flags frauds by how different the values are from normal transactions.
- **Predictive Analytics**: Use statistical models or machine learning models to generate a score for transactions or sign-up attempts. Depending on the type of model used, the interpretability varies.

4 Data Quality Issues and Causes

4.1 By the Nature of Data

4.1.1 Imbalanced Data

When we build classification models to predict whether a customer or transaction is fraudulent or not, having very few observations of frauds versus the majority are non-frauds creates the class imbalance problem. In the real world, fraudulent transactions and signups are the minority. According to research from the Federal Reserve Bank of Kansas City (Hayashi 2019), credit card fraud rate in the US for no chip credit cards is about 16 basis points (0.16%) in 2016, and that of chip credit cards is about 12 basis points (0.12%).

This is because people who want to commit frauds are the minority of the population. Based on the theory of "Fraud Triangle", someone needs to have the motivation, rationalization, and opportunity to commit frauds (National Whistleblower Center 2023). Besides, financial institutions actively employ fraud prevention methods, so even though there could be fraud attempts, the successful ones are quite few. When we need a clear label for fraud, we only consider those attempts that got past the prevention methods and eventually confirmed to be fraud either by disputes or chargebacks.

Having imbalanced data could cause the model training to be biased towards predicting the majority class. For example, if only 1% of the samples are real frauds, then if a model predicts 100% of the samples to be frauds, it still has 99% accuracy. Besides, even with methods like oversampling, it could cause overfitting to the limited number of fraud samples.

4.1.2 Wrong Labels

As the target variable, fraud or non-fraud as a binary value is usually sourced from chargebacks or customer disputes. Even though there are certain patterns to infer fraud vs non-fraud, sometimes drawing conclusions for the cases still rely on human investigation. For example, when a customer disputed the transaction as "unauthorized purchase", it could be that the customer is telling the truth, and it's a third-party fraud, but it could also be that the customer lied about it, and in this case it's a first-party fraud. The patterns of these two types of frauds are different and mis-labelling the records leads to worse model performance.

Besides, mis-labeling could also come from manual entry errors. For example, if the agent handling the dispute closed the case with reason "wrong merchandise" instead of "unauthorized purchase", then it's considered non-fraud in subsequent analyses, while it actually should be fraud.

4.1.3 Incomplete Data

Credit card signup data has a lot of personal information, so it's subject to strict regulations in most regions such as the General Data Protection Regulation (GDPR) in the European Union. In the case of GDPR, consumers have the right to opt out of certain data collection, which could leave fewer attributes for companies to use for analytics, including fraud detection. Individuals also have the right to have their data deleted, limiting the types of data available.

4.2 By Procedures

4.2.1 Latency

During the data ingestion phase, there could be large backlogs of tasks to be run, creating lag between data being generated and data being loaded into storage. This lag could create a gap between when a fraud attack happens and when teams are able to get the relevant data to conduct root cause analysis. Based on industry case studies, this gap could be 1 hour (Materialize 2024) up to several days. Even just 1 hour of lag could cost companies thousands to tens of thousands of dollars.

4.2.2 Wrong Calculations

Transformation processing is using filters, operations, and functions to turn the source data into desired format and metrics. When Analysts or Data Engineers write the transformation logic, they could make mistakes, resulting in wrong calculations. For example, when joining different tables together to form new tables, missing a join key could create duplicate records. If metrics are calculated based on duplicated records, it could deviate from the actual metrics.

4.2.3 Data Silos

Credit card issuers collect data from different sources, and the integration and ingestion processes could be owned by different teams. Without a proper Data Catalog, teams don't always know what each other is processing. This creates data silos, which means different sets of data exist, but are not being used together.

For example, the Marketing team could send out a large campaign, causing credit card sign-up applications to significantly increase in a short period of time. If the fraud team is not informed of this campaign, they would have a hard time identifying whether this application volume spike is from a fraud attack or because of a marketing campaign, which mainly brings in non-fraud customers. This causes hours to days of labor that could have been saved if there's a clearer visibility of datasets.

4.3 By Nature of the Task

4.3.1 Afterthought of the Business

Because data quality assurance is usually considered an operational or back-office function, it does not get as many resources as revenue-generating functions, such as Product or Sales. When data quality is high, it does not get noticed because things are working as expected. However, when the data quality drops and causes incidents, that's when it gets visibility.

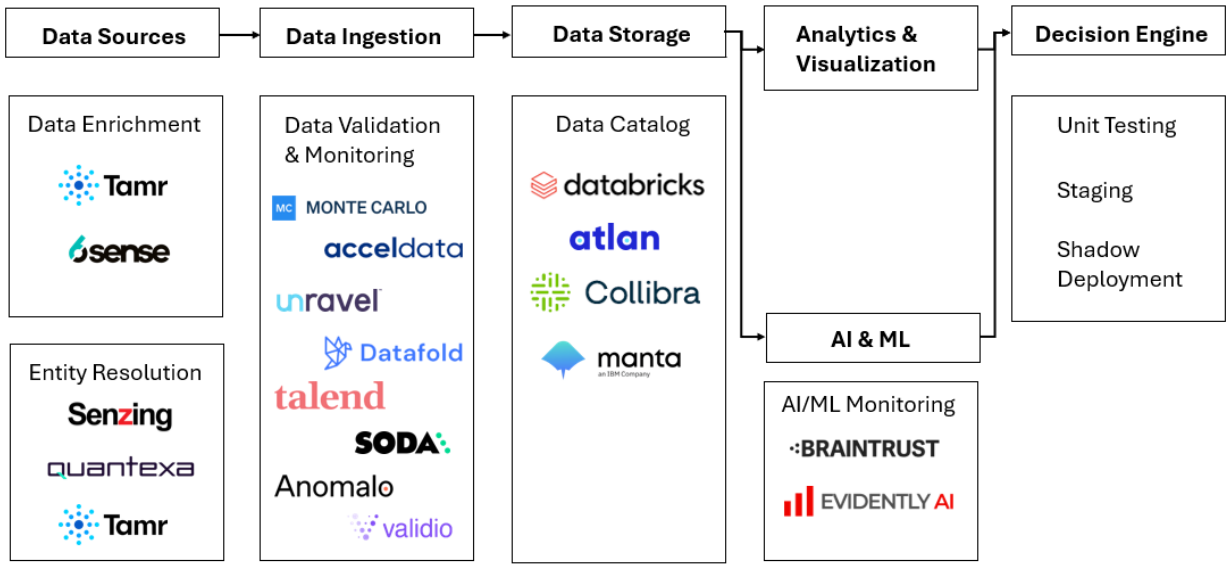
4.3.2 Data Quality Assurance could be Tedious

Because data quality assurance involves a lot of ad hoc analysis to find out the root cause, it could be tedious. It also does not have very tangible output, so it doesn't usually provide much sense of fulfillment. Therefore, it's hard to attract talents to work in this field, even though it's an important function. When there's a shortage of talents to work on it, there aren't sophisticated enough checks on data quality.

5 Data Quality Assurance Solutions

5.1 Solutions Diagram

This diagram maps the solutions and representing products / companies to the Fraud Prevention data stack.



5.2 Data Validation

Data validation is the process of validating that the data has the expected data type, range, and format etc. In Fraud Prevention, common validation checks include:

- **Data type**: Customer names should be in string type, and date of birth should be in date type.
- **Range**: Transaction date should be after the card was created; date of birth doesn't exceed biological constraints (e.g. not possible to have a customer who is 200 years old)
- **List of values**: Dispute reasons need to conform industry standards, for example, "unauthorized purchase", "wrong merchandise".
- **Format**: National Identification Number should have the same format for the same country.

Data validation could be built into the user interface for data entry. For example, when customers enter their street address, have them select from USPS verified addresses. Internally, the company can build standardized labels for fraud investigators to choose from when they close the cases, instead of having free text entry as labels. Besides, there should be sampling on the labelled cases, to verify that the agents didn't tag the cases randomly to sacrifice the quality for speed.

5.3 Entity Resolution

Entity Resolution matches different records that are referring to the same thing and resolve them into the same entity. For example, a customer could have a legal name and preferred name. Using entity resolution methods, we can link these two different objects together and get a more holistic view than without it. This is especially helpful for fraud linking analysis, which leverages graph structure, because it splits the same entity into different nodes when the source records have different attributes.

Therefore, introducing Entity Resolution to fraud analytics improves data quality and creates positive business impact. According to Entity Resolution provider Quantexa, its entity resolution applied in fraud

prevention achieved an 80% reduction in investigation time and a 75% reduction of false positives (Quantexa 2024).

5.4 Data Enrichment

Data Enrichment is the process of integrating external data sources and combining it with internal data, to add missing details of the entity. For example, customers may only submit very few essential fields, but credit card issuers can query vendors or bureaus to get more information about the customer. This alleviates the problem of missing data.

5.5 Rule-based Monitoring

A key element in data quality assurance is establishing metrics and monitoring them. There are some commonly used metrics across different domains like percentage of nulls in a column, percentage of anomalies in a column, number of duplicated records. Business logic determines what metrics should be used, for example, what is considered an anomaly – shall we use 2 standard deviations away or 3 standard deviations as the threshold. Rule-based checks could be done at each destination of data loading. However, because this approach is not very scalable, many organizations only use it for critical datasets. Discover observes that it takes about 40 hours for a Data Analyst to design a quality check for a column (Jaganathan 2023).

5.6 Data Catalog

Data Catalog is a metadata repository of datasets in the organization which includes but not limited to Data Dictionary, Data Lineage, and Data Ownership. It is an important tool to manage datasets and improve data quality.

- A **data dictionary** has field names and explanation of the dataset, which helps users understand how to use it. This helps identify misusage of fields and helps define validation rules.
- **Data lineage** shows the upstream and downstream datasets, which helps assess impact of data issues for downstream processes, and trace back the root cause through the upstream processes.
- Showing **data ownership** helps quickly resolve data issues when they are discovered.

5.7 ML-Powered Monitoring

Machine Learning (ML) techniques automatically discover patterns and train for the best parameters for predictions. Using ML to power data quality monitoring frees Data Analysts from manually defining rules. For example, using Unsupervised Learning techniques to conduct data quality checks helps discover problems humans cannot think of. For example, a human analyst could define intervals to use for alerting anomalies, but it's hard to detect seasonality in data.

Typical models used in data quality monitoring include:

- **Tree based models**: Easy for interpretation and implementation; useful for detecting categorical inconsistencies.
- **Support Vector Machines (SVM)**: Can be used for separating high quality and lower quality data points; scalable.
- **K-Nearest Neighbors (K-NN)**: Used for imputing missing values.
- **K-Means**: Group similar data points together and identify anomalies that don't belong to larger groups.

6 Conclusion and Future Improvements

This paper proposed solutions of data issues in Credit Card Fraud Prevention by looking at the object of the issues, processes, and tools involved.

Firstly, it highlighted why data quality is essential in the problem space. Then, it summarized the types of data used in this space: customer information, transactions, behavioral data, external data sources like bureau data and social media data. The diversity of data sources contributes to the complexity of data management and problem of data silos. Afterwards, it illustrated the commonly used data stack in Credit Card Fraud Prevention, using a diagram to show the relationships of data sources, ingestion pipelines, storage, analytics, visualization, ML/AI, and decisioning. It further discussed the causes of data issues through different dimensions, including the nature of data, procedures, and the nature of the tasks. Lastly, it suggested solutions including data validation, entity resolution, data enrichment, data monitoring, and data catalogs.

For future work, we will use real-world case studies of FinTech companies to show the impact of applying the best practices of data quality assurance, and how that prevents losses. Besides, as Artificial Intelligence possesses the center of the stage in technology, we would discuss how AI is used in Credit Card Fraud Prevention and how AI observability should be conducted in this field.

References

Atlan. n.d. "6 Reasons Why Data Quality Needs a Data Catalog." *Atlan.* https://atlan.com/data-quality-with-data-catalog/#:~:text=A%20good%20data%20catalog%20can,can%20inform%20data%20quality%20measures.

Federal Trade Commission. 2024. "As Nationwide Fraud Losses Top $10 Billion in 2023, FTC Steps Up Efforts to Protect the Public." *Federal Trade Commission.* 2 9. https://www.ftc.gov/news-events/news/press-releases/2024/02/nationwide-fraud-losses-top-10-billion-2023-ftc-steps-efforts-protect-public.

Hayashi, Fumiko. 2019. "Payment Card Fraud Rates in the United States Relative to Other Countries after Migrating to Chip Cards." *Econimic Review*, 12. Accessed July 2024. https://www.kansascityfed.org/Economic%20Review/documents/681/Payment_Card_Fraud_Rates_in_the_United_States_Relative_to_Other_Countries_since_Migrat.pdf.

Jaganathan, Prakash. 2023. "Discover's Approach to Scaling Enterprise Data Quality Monitoring." *Anomalo.* 10 17. Accessed 2024. https://www.anomalo.com/case-study/discovers-approach-to-scaling-enterprise-data-quality-monitoring/.

Materialize. 2024. "Real-Time Fraud Detection: Analytical vs. Operational Data Warehouses." *Materialize.* 3 7. Accessed 7 9, 2024. https://materialize.com/blog/fraud-detection-latency-accuracy/.

National Whistleblower Center. 2023. *The Fraud Triangle.* Accessed 2024. https://www.whistleblowers.org/fraud-triangle/#:~:text=According%20to%20Albrecht%2C%20the%20fraud,being%20inconsistent%20with%20one's%20values.%E2%80%9D.

Quantexa. 2024. "Create context to counter the rise of fraud." *Quantexa.* Accessed 7 10, 2024. https://www.quantexa.com/solutions/fraud/#challenge.

SAS. 2021. *State of Insurance Fraud Technology Study.* SAS.

So, Kenn, and Ben Lorica. 2020. *Gradient Flow.* https://gradientflow.com/data-quality-unpacked/.

Striim. 2024. "Data Warehouse vs. Data Lake vs. Data Lakehouse: An Overview of Three Cloud Data Storage Patterns." *Striim.* https://www.striim.com/blog/data-warehouse-vs-data-lake-vs-data-lakehouse-an-overview/.

Worldpay. 2024. *The Global Payments Report.* Worldpay.