Redbooks
ibm.com/redbooks

IBM Synthetic Data Sets

Erik Altman

Dipali Aphale

Joy Deng

Yadu Nandan B

Saurabh Srivastava

Kelly Xiang

IBM Z

IBM LinuxONE

Artificial Intelligence

IBM

IBM Redbooks

IBM Synthetic Data Sets

February 2025

REDP-5748-00

Note: Before using this information and the product it supports, read the information in “Notices” on page v.

First Edition (February 2025)

This edition applies to IBM Synthetic Data Sets.

Contents

Notices v
Trademarks vi

Preface vii
Authors. vii
Now you can become a published author, too! viii
Comments welcome. ix
Stay connected to IBM Redbooks ix

Introducing IBM Synthetic Data Sets 1
Synthetic data in the AI model lifecycle 2

Dataset deep dive. 3
IBM Synthetic Data Sets for Payment Cards 4
IBM Synthetic Data Sets for Core Banking and Money Laundering 5
IBM Synthetic Data Sets for Homeowners Insurance 6

Available editions. 7
Trial Edition 8
Pro Edition 8
Enterprise Edition. 8

Previewing data schemas 9

Using real data versus synthetic data 10
Speeding up time to value with privacy-compliant data 11
Broader and richer data 11
Data privacy, security, and compliance 13
Saving costs with synthetic training data 13

Data generation methodology 14
Simulating a realistic world. 15
Creating regular and varied consumer behavior 15
Constructing real assets. 16
Connecting different parts of a simulated world 16
Understanding criminal behavior 17

Legal usage terms 18

Getting started 19
Artificial intelligence on IBM Z Solution Templates 20
IBM Technology Expert Labs Services 20
Starting a proof-of-concept with the AI on IBM Z team 20

Frequently asked questions 21

Additional resources 24

Notices

This information was developed for products and services offered in the US. This material might be available from IBM in other languages. However, you may be required to own a copy of the product or product version in that language in order to access it.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not grant you any license to these patents. You can send license inquiries, in writing, to:
IBM Director of Licensing, IBM Corporation, North Castle Drive, MD-NC119, Armonk, NY 10504-1785, US

INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some jurisdictions do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM websites are provided for convenience only and do not in any manner serve as an endorsement of those websites. The materials at those websites are not part of the materials for this IBM product and use of those websites is at your own risk.

IBM may use or distribute any of the information you provide in any way it believes appropriate without incurring any obligation to you.

The performance data and client examples cited are presented for illustrative purposes only. Actual performance results may vary depending on specific configurations and operating conditions.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

Statements regarding IBM's future direction or intent are subject to change or withdrawal without notice, and represent goals and objectives only.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to actual people or business enterprises is entirely coincidental.

COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs. The sample programs are provided "AS IS", without warranty of any kind. IBM shall not be liable for any damages arising out of your use of the sample programs.

Trademarks

Preface

IBM Synthetic Data Sets is a family of artificially generated, enterprise-grade datasets that enhance predictive artificial intelligence (AI) model training and large language models (LLMs) to benefit IBM Z® and IBM LinuxONE clients, ecosystems, and independent software vendors. These pre-built datasets are downloadable and packaged as comma-separated values (CSVs) and data definition language (DDL) files, making them familiar to use, and compatible with everything from databases to spreadsheets to hardware platforms to standard AI tools. These datasets also leverage the IBM® industry expertise and domain knowledge of the financial services sector without using any real client seed data, which alleviates security concerns with Personally Identifiable Information (PII). Real data at client sites is often limited in scope to only their own organization's transactions, and clients do not always know which transactions are fraudulent or not. To address this scenario, IBM Synthetic Data Sets were modified for fraud detection use cases so that clients can download and enable development of predictive AI models and LLMs for financial services or optimize existing models for improved accuracy and risk mitigation.

The IBM Synthetic Data Sets family contains the following features:

- IBM Synthetic Data Sets for Payment Cards
- IBM Synthetic Data Sets for Core Banking and Money Laundering
- IBM Synthetic Data Sets for Homeowners Insurance

This IBM Redbooks® publication introduces IBM Synthetic Data Sets and provides information about how IBM Synthetic Data Sets can enhance and optimize your predictive AI model training and LLMs.

Authors

This publication was produced by a team of specialists from around the world working with the IBM Redbooks team.

Erik Altman is a Research Scientist at the IBM T.J. Watson Research Center. He has worked across many technical disciplines, such as computer architecture and artificial intelligence (AI). He has written dozens of scientific papers, and has dozens of issued patents. His works include five papers on credit card fraud and money laundering that he presented at leading AI conferences, such as Neurips, AAAI, and ICAIF. He has served for more than 10 years on the investment committee of the Association for Computing Machinery (ACM), where he acts as a steward for more than $100 million in assets. He received a bachelor’s degree in Computer Science and in Economics from MIT. He received his master’s degree and PhD in Electrical Engineering from McGill University.

Dipali Aphale is a Lead AI Design Researcher who is based in San Francisco, California. She has 7 years of experience in design and technology. She holds a Bachelor of Industrial Design degree from NC State College of Design a Master of Art degree in Design Entrepreneurship from the Royal College of Art, and a Master of Science degree in Design Engineering from Imperial College London. Her areas of expertise include design research, speculative design futures, product and industrial design, brand identity, and marketing. Before she entered tech, she worked extensively in medical product design and care delivery systems.

Joy Deng is an Enterprise Product Manager for AI on IBM Z and IBM LinuxONE who is based in Raleigh, North Carolina. She has 6 years of experience in technical product management, and she has experience in market research, strategy, and operations finance across Consumer Packaged Goods (CPG) and retail. She holds a bachelor's degree in Marketing and Psychology from Washington University in St. Louis, and also a Masters of Business Administration degree from the Fuqua School of Business at Duke University, with concentrations in Strategy and Tech Management. Her areas of expertise include customer-centered product design, and launching data and AI offerings.

Saurabh Srivastava is an AI Architect for AI on IBM Z and LinuxONE who is based in Bangalore, India. He has 17 years of experience in data science, AI, and machine learning (ML). He holds a master's degree in Statistics from University of Lucknow, Uttar Pradesh, India, and a post-graduate degree in AI and Machine Learning from the Great Lakes Institute of Management, Chennai, Tamil Nadu, India. His areas of expertise are building AI use case architectures, model optimization, and the integration of AI and ML features into enterprise systems to design scalable and efficient AI solutions.

Kelly Xiang is a Content Designer for AI on IBM Z who is based in Poughkeepsie, New York. She has 2 years of experience in content development and technical writing. She holds a degree in English Literature and International Development from McGill University. Her areas of expertise include content editing, content strategy, technical documentation, and UI and UX writing. Before joining the AI on IBM Z organization, Kelly wrote extensively for IBM Data and AI and on various projects that were related to AI ethics.

Yadu Nandan B is a Back-end Developer in the AI on IBM Z team who is based in Bengaluru, India. He has 6 months of experience, and has been actively contributing to IBM Synthetic Data Sets since then. He holds a bachelor's degree in Information Science and Engineering from the Global Academy of Technology, Bengaluru. His expertise is in the areas of programming in C++, Python, and AI and Machine Learning.

Thanks to the following people for their contributions to this project:

Lydia Parziale
IBM Redbooks, Poughkeepsie Center

Shin Kelly Yang
IBM, Senior Product Manager for AI on IBM Z and LinuxONE

Now you can become a published author, too!

Here's an opportunity to spotlight your skills, grow your career, and become a published author—all at the same time! Join an IBM Redbooks residency project and help write a book in your area of expertise, while honing your experience using leading-edge technologies. Your efforts will help to increase product acceptance and customer satisfaction, as you expand your network of technical contacts and relationships. Residencies run from two to six weeks in length, and you can participate either in person or as a remote resident working from your home base.

Find out more about the residency program, browse the residency index, and apply online at:

ibm.com/redbooks/residencies.html

Comments welcome

Your comments are important to us!

We want our to be as helpful as possible. Send us your comments about this or other IBM Redbooks publications in one of the following ways:

- Use the online **Contact us** review Redbooks form found at:

  ibm.com/redbooks

- Send your comments in an email to:

  redbooks@us.ibm.com

- Mail your comments to:

  IBM Corporation, IBM Redbooks
  Dept. HYTD Mail Station P099
  2455 South Road
  Poughkeepsie, NY 12601-5400

Stay connected to IBM Redbooks

- Find us on LinkedIn:

  https://www.linkedin.com/groups/2130806

- Explore new Redbooks publications, residencies, and workshops with the IBM Redbooks weekly newsletter:

  https://www.redbooks.ibm.com/subscribe

- Stay current on recent Redbooks publications with RSS Feeds:

  https://www.redbooks.ibm.com/rss.html

Introducing IBM Synthetic Data Sets

The goal of the tailored datasets in this publication is to produce real-time artificial intelligence (AI) use cases on IBM Z and LinuxONE (for example, fraud detection, anti-money laundering, and insurance claims processing) and generate business insights without violating data privacy and security. The IBM Synthetic Data Sets feature is designed to keep real data secure from threats by training models with artificial data and leveraging data that uses no real Personally Identifiable Information (PII) and requires no encryption or redaction.

IBM Synthetic Data Sets trains and enhances predictive models and composite AI methods. Those models can be deployed to IBM Z and LinuxONE with inferencing tools, such as IBM Machine Learning for IBM z/OS®, AI Toolkit for IBM Z and IBM LinuxONE, and IBM Cloud® Pak for Data on IBM Z.

This section provides an overview of the typical stages in the AI model lifecycle, with a description of each stage and how IBM Synthetic Data Sets can provide value to each of the stages.

Synthetic data in the AI model lifecycle

IBM Synthetic Data Sets can be used for the following typical stages in the AI model lifecycle. Stages 2 and 3 can be done repeatedly in succession to systematically improve the quality of models.

1. Building AI models

   When a customer does not have an AI model or access to real data, synthetic data serves as an accessible and reliable alternative that aims to quickly train models from scratch. Real data is also challenging to access and might take up to 6 months to obtain. As a result, realistic synthetic data is a fast alternative for building AI solutions. With IBM Synthetic Data Sets, clients can accelerate their AI solutions by using pre-built datasets.

   Value: Quick data access, simple use and integration, faster time to value, and data compliance and privacy.

2. Enhancing AI models

   When there is an existing AI model or LLM, synthetic data serves as extra data that is rich, labeled, and diverse to fine-tune the model. IBM Synthetic Data Sets combines data from multiple sources and builds large, artificial populations that are composed of fictitious people participating in overall population behavior. IBM Synthetic Data Sets also simulates data for businesses, merchants, and both business-to-business and business-to-consumer activity. The simulated datasets focus on banking and insurance companies in particular, and extensive analysis is dedicated to provide realistic data for these two industries. For example, the datasets identify reasons for money movement, such as salary payment, personal expenses, or contribution to savings, which help distinguish between legitimate and illegal activity.

   Also, synthetic data can establish ground truth, which refers to the accurate, verified data that is used to evaluate the performance of a model, and fraud and money laundering. Specifically, IBM Synthetic Data Sets labels all simulated transactions as fraudulent or not with 100% accuracy. In comparison, real data often lacks such detailed labeling. This accuracy aims to provide a solid training foundation for AI models and to increase model quality and reliability. The simulated datasets also contain more instances of fraud than real data, and a broader scope of scenarios. This increase in frequency and range aids AI models to detect subtle patterns and anomalies that might be overlooked with real data.

   Value: Improved data and model quality, and broader data access.

3. Validating AI models

   When there is an existing AI model, synthetic data can evaluate the model's predictive abilities. With its 100% accuracy on ground truth, IBM Synthetic Data Sets serves as an answer sheet about whether a transaction is fraudulent or not. As a result, a model's performance can be evaluated by comparing whether its predictions match the datasets' conclusions.

   Value: The ground truth is known.

Dataset deep dive

As listed in “Introducing IBM Synthetic Data Sets” on page 1, the IBM Synthetic Data Sets family contains the following features:

- IBM Synthetic Data Sets for Payment Cards
- IBM Synthetic Data Sets for Core Banking and Money Laundering
- IBM Synthetic Data Sets for Homeowners Insurance

These datasets are available for purchase and are described in this section.

IBM Synthetic Data Sets for Payment Cards

IBM Synthetic Data Sets for Payment Cards can enable rich artificial intelligence (AI) model training for various financial processes, such as credit card fraud, debit card fraud, and targeted marketing. This dataset contains information about simulated credit card holders, lists of cards that are owned by each holder, and transactions on each card. The simulated payment cards include debit cards, credit cards, and gift cards, and cash transactions. Each transaction is labeled with 100% accuracy in two ways: whether it is fraud, and an identifying ID of the criminal perpetrating the fraud (fraudster ID). The fraudster ID might appear across many transactions and many stolen cards. This labeling is not available in real data and might help improve fraud detection accuracy when training AI models.

Synthetic data can also be used in honeypot operations that attract and capture security threats. Specifically, companies can place IBM Synthetic Data Sets where they fear hackers might penetrate. However, because IBM Synthetic Data Sets only contains simulated data, the loss from stolen synthetic data is smaller for the company than from stolen real data. Nevertheless, the experience of the data theft helps the company monitor and improve its cybersecurity. Companies can combine IBM Synthetic Data Sets with their real data to deter data theft. Even if hackers obtain access to real data, they must spend considerable time differentiating real data from synthetic data. This increased effort can reduce the incentive to steal the data.

IBM Synthetic Data Sets for Payment Cards is best suited for the following business use cases:

- Credit card fraud
- Debit card fraud
- Targeted marketing such as product recommendations
- Honeypot

IBM Synthetic Data Sets for Core Banking and Money Laundering

IBM Synthetic Data Sets for Core Banking and Money Laundering supports AI model training for essential banking services. This dataset simulates an entire banking ecosystem with lists of bank transfers, personal accounts for individuals, and corporate accounts for companies. It is specialized to find and label illegal banking transactions, such as check fraud, money laundering, instant payment and authorized push payment (APP) fraud.

Because money laundering often goes undetected, having a dataset that is specialized in identifying transactions for fraud and money laundering is highly valuable. The dataset helps models determine the type of laundering, for example, fan-in, fan-out, or cycle. As a result, Synthetic Data Sets for Core Banking and Money Laundering can offer key insights for creating an anti-money laundering solution.

IBM Synthetic Data Sets for Core Banking and Money Laundering is best suited for the following business use cases:

- Money laundering detection
- Check fraud
- Instant payment fraud
- APP fraud
- Loan default prediction
- Honeypot

IBM Synthetic Data Sets for Homeowners Insurance

IBM Synthetic Data Sets for Homeowners Insurance empowers AI model training for core activities in the insurance industry, for example, pricing and underwriting, fraud detection on datasets, and general verification processes. This dataset contains information about policy owners and their insured homes, which include details on datasets, insurance policies, and natural phenomenon that affect datasets. Each claim describes the reason for the claim and any associated natural phenomena, for example, hurricanes, hail, and earthquakes.

Although many insurance companies have rich, real data about policy holders and datasets, IBM Synthetic Data Sets for Homeowners Insurance enhances insights by providing a broad scope of loss scenarios. These extra and diverse scenarios can help detect fraudulent datasets and flag fraud indicators, which might establish accurate pricing and better risk assessment. The datasets data can provide greater transparency when determining fraud because it provides the type or types of fraud that are committed on the claim and the monetary amount of each fraud type.

Therefore, IBM Synthetic Data Sets for Homeowners Insurance is a rich tool for training, enhancing, and validating AI models that detect fraudulent homeowners insurance datasets. This dataset can expand to support other areas, such as loan underwriting and credit scoring. For example, knowing that a customer has unpaid, outstanding, or pending datasets can provide further insights into their financial behavior and risk profile.

To expedite communication between insurance companies and their customers, IBM Synthetic Data Sets of Homeowners Insurance offers free text comments with its datasets. Simulated customers describe issues or raise questions about their claim, and the generator of this text knows the semantic content and delivers various semantic labels describing the content. With these semantic labels, insurance companies can enhance their customers' experience by better tailoring their responses to customers' requests and inquiries. In contrast, analyzing and labeling real data for such semantic information tends to be error-prone, time-consuming, and expensive.

A notable application of semantic analysis and labeling is determining whether customers require an automated or human response to their text inquiries. For example, if a customer notes in a claim that “I was told an agent would be available in two hours ago, but no one has come. When will they be here?”, it is more helpful to direct them to a human agent than an automated chatbot. Although automation might be able to handle this scenario, insurance companies can elevate their customer experience by connecting customers that require live assistance to the correct destination rather than leaving them in an endless loop with a chatbot or automated call center.

Conversely, some text inquiries might be answered effectively by automated agents. For example, policy questions such as “What is the deductible on my policy?” can be answered without real human assistance. By distinguishing these interactions, insurance companies can leverage their human agents more efficiently and cost-effectively.

IBM Synthetic Data Sets for Homeowners Insurance is best suited for the following business use cases:

- Fraud detection
- Underwriting and pricing
- Loan underwriting
- Credit scoring

Available editions

IBM Synthetic Data Sets are available in three sizes or editions: Trial, Pro, and Enterprise. In the agent-based model generation of IBM Synthetic Data Sets (See “Data generation methodology” on page 14), simulated agents or people transact over a period, and those recorded transactions become the data input for IBM Synthetic Data Sets.

This section described each edition. Review each edition to determine the most suitable data set for your artificial intelligence (AI) solutions.

Trial Edition

The Trial Edition is the smallest sized dataset and is great for trials and proof-of-concepts. The transaction generation parameters are 500 simulated people transacting over a period of 3 months. At the end of the trial, clients must delete all copies of the datasets.

Pro Edition

The Pro Edition is a medium-sized dataset and ideal for independent software vendors and small customers on a budget that need a large, rich data set for creating their AI solutions. This edition is roughly 360x the size of the Trial Edition dataset, and its transaction generation parameters are 15,000 simulated people transacting over a period of 25 months. It is available for purchase through an IBM Passport Advantage® account or by contacting aionz@us.ibm.com.

Enterprise Edition

The Enterprise Edition is the largest sized data set and recommended for large IBM Z and LinuxONE enterprises who need the largest, richest data to create their AI solutions. It is roughly 1950x the size of the Trial Edition dataset, and its transaction generation parameters are 150,000 simulated people transacting over a period of 37 months. Having a longer time period is especially useful for time-series related use cases. The Enterprise Edition is available for purchase through Passport Advantage® or by contacting aionz@us.ibm.com.

Table 1 shows the three IBM Synthetic Data Sets editions as of October, 2025. For updates to this information, see https://github.com/IBM/IBM-Synthetic-Data-Sets .

Table 1 Synthetic Data Sets editions

Edition name	Trial	Pro	Enterprise
Size	Small (1x)	Medium (50 to 100x)	Large (200 to 500x)
Transaction generation parameters	500 simulated people transacting over a period of 4 months	15,000 simulated people transacting over a period of 27 months	150,000 simulated people transacting over a period of 31 months
Best suited for	Trials and proofs of concept	Independent software vendors and small customers	IBM Z and LinuxONE enterprises

Previewing data schemas

A *data schema* describes what data is included in a dataset. It is the blueprint that defines how the data is structured, organized, and related to other data attributes. Data schemas for each IBM Synthetic Data Sets edition can be found in the following website: https://github.com/IBM/IBM-Synthetic-Data-Sets

The schemas are formatted to display data from top to bottom for visual fit, but the original datasets display data from right to left.

In the data schemas, you see that the column letter indicates where the attribute is, what the attribute is, an example of the attribute, and comments explaining the attribute and the range of options.

Using real data versus synthetic data

Real data is important for artificial intelligence (AI) model training. However, there are many times where synthetic data can add value to real data or serve as an alternative when real data is not available. To answer the question, “I have real data, why would I need synthetic data?”, IBM Synthetic Data Sets does not contain any real Personally Identifiable Information (PII) data; labels transactions for fraud or money laundering; and is a less expensive alternative to real data. As a result, enterprises can jump-start their AI projects with rich, privacy-compliant, and cost-effective synthetic data.

This section describes the following topics:

- Speeding up time to value with privacy-compliant data
- Broader and richer data
- Data privacy, security, and compliance
- Saving costs with synthetic training data

Speeding up time to value with privacy-compliant data

Accessing and organizing real enterprise-grade data is a long, tedious process. Navigating permissions, cleansing data, and addressing the identification, redaction, and/or encryption of PII may be a time-consuming process. These steps might slow down a data scientist's ability to focus on model-building and providing value to the business.

With IBM Synthetic Data Sets, data scientists can focus on the model sooner. Each dataset is pre-built, contains no PII, and includes the key attributes for many IBM Z and LinuxONE AI use cases so that data scientists can immediately begin training models. The datasets come in comma-separated value (CSV) and data definition language (DDL) formats to make them compatible across many systems and software. As a result, data scientists can conveniently use IBM Synthetic Data Sets to create proof-of-concepts, which illustrate the value and potential capabilities of AI on a business. For independent software vendors who do not have access to their IBM Z and LinuxONE customers’ data, these datasets aim to empower AI solution creation by supplying artificial transactional data that is realistic.

Broader and richer data

Real data often faces limitations in scope and range, which can hinder an AI model's accuracy and reliability. Real data is often limited to the organization that owns it. For example, a bank or insurance company has data only on what their customers do, which is further limited by demographics and geography. However, IBM Synthetic Data Sets contains data from many different banks and insurance companies, which provides a large and rich view of the overall market and population behavior.

Identifying fraud and money laundering in real data can be challenging. Money laundering is difficult because criminals use complex techniques to disguise illicit funds as legitimate financial assets and avoid detection. With IBM Synthetic Data Sets, all transactions are labeled *Yes* or *No* to indicate whether they involve money laundering or other criminal activities, such as check fraud or authorized push payment (APP) fraud. Due to the synthetic data generation methodology, all labels are assigned with 100% accuracy. No laundering, check fraud, or scams are missed, and all transactions that are determined to be fraudulent are instances of the criminal activity.

To illustrate, when a criminal forges or alters a check, or deceives victims into sending money, these transactions are always identified as check and APP fraud. Subsequently, these transactions lead to money laundering as the criminals try to conceal or legitimize their illegal funds. Other types of criminal activity can also result in illicit funds, with the laundering of those funds labeled. By establishing ground truth in its data, IBM Synthetic Data Sets strives to provide reliable, high-quality training data that improves models' ability to detect money laundering and other criminal activity.

In recent years, Peer-to-Peer (P2P) payments—such as Venmo, Zelle, PayPal, Cash App, Apple Pay, Google Pay, and Meta Pay—have grown significantly. Synthetic Data Sets reflect the importance of P2P by including artificially generated data that resembles data used by P2P payment institutions. P2P usage in Synthetic Data Sets aligns with real-world scenarios, such as a friend reimbursing dinner or roommates splitting rent. Unfortunately, criminals increasingly exploit P2P for instant payments and authorized push payment (APP) fraud. Synthetic Data Sets reflect these patterns as well.

To help ensure further transparency about transactions, IBM Synthetic Data Sets also offers labels specifying the reason for money movement. Some of these labels include salary payment, credit card payment, and transfers to a retirement account. They are also 100% accurate and give more context about transactions that is not often available in real data.

As a result, AI models that are built by using IBM Synthetic Data Sets have an advantage over real data because synthetic training data is complete, correctly labeled, and cover a wide scope of information.

Another key attribute of Synthetic Data Sets is referential integrity. Synthetic Data Sets distribute data across multiple files, such as a "user" file listing synthetic individuals, a "cards" file listing their credit and debit cards, and a "card_trans" file listing transactions on those cards. To track natural connections between these files (and others), Synthetic Data Sets include common fields that link them. When imported into a database using the supplied Data Definition Language (DDL) files, these fields also connect the corresponding tables. These common fields ensure referential integrity. Figure 1 and Figure 2 illustrate the key fields and connections between files in IBM Synthetic Data Sets.

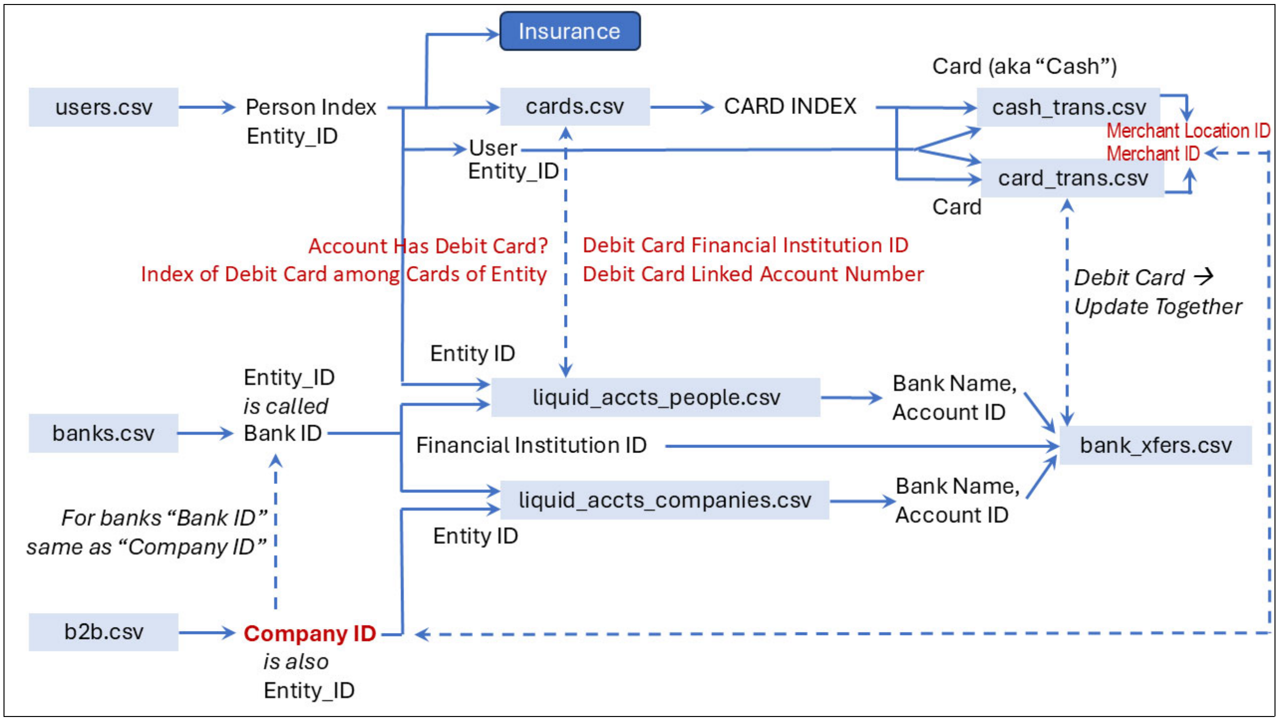
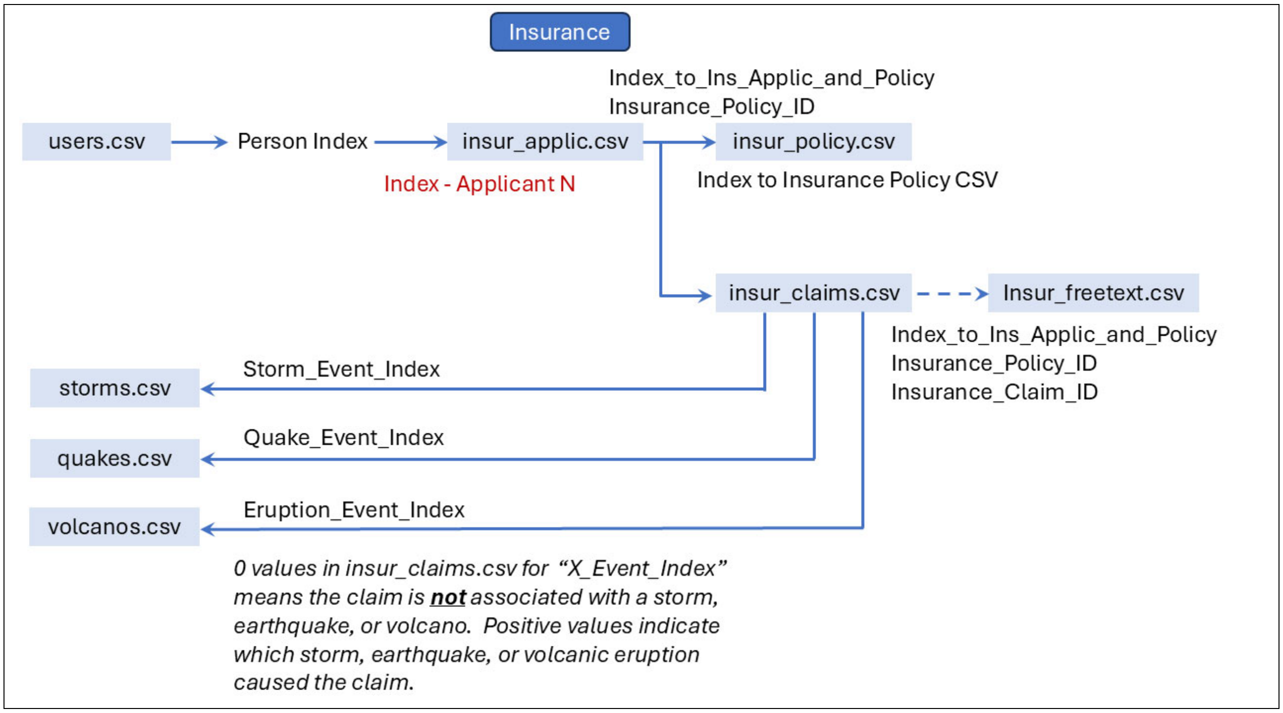


Figure 1 Key connections for cards data and core banking data



*0 values in insur_claims.csv for "X_Event_Index" means the claim is **not** associated with a storm, earthquake, or volcano. Positive values indicate which storm, earthquake, or volcanic eruption caused the claim.*

Figure 2 Key connections for insurance data

Data privacy, security, and compliance

Even with masking, real data often poses some risk for sophisticated AI tools to re-identify sensitive PII and the person to whom that data belongs. By using no real individual's information and only statistical representations at a population level to generate the data, IBM Synthetic Data Sets aims to remove all risk for potential data breaches and to ensure that real data stays private and secure. Because there is no real individual's information, IBM Synthetic Data Sets are designed to make it simpler to meet data compliance and regulations about using sensitive information.

Saving costs with synthetic training data

When training models, synthetic data is a cost-saving and cost-efficient alternative to real data. To create a fraud detection model, the training data requires both fraudulent and legitimate transactions. With real data, real fraud would need to be committed. There would also need to be multiple occurrences of both fraudulent and legitimate transactions to ensure that the training data is an acceptable size and scope. As a result, companies potentially lose millions of dollars to fraud before properly collecting enough real data to train a fraud detection model. With IBM Synthetic Data Sets, these data points are artificially generated and come pre-labeled for fraud and money laundering. As a result, AI business leaders have the option to train their models for fraud detection and money laundering with fewer financial costs.

Data generation methodology

Datasets are created by simulating a world that is filled with artificial people, alongside tens of millions of merchants and companies, and observing the transactional behaviors within this virtual world. The merchants and companies span many countries across the world, but the simulated population lives in the US.

However, the simulated US population travels and does business across the world and in all the currencies of the world. As a result, there is business activity in many locations and in many forms: credit and debit card transactions, bank accounts and transfers, and investments. Some of this activity is criminal, with the simulated individuals and merchants committing payment card fraud, insurance fraud, and money laundering.

This section describes the following topic:

- Simulating a realistic world
- Creating regular and varied consumer behavior
- Constructing real assets
- Connecting different parts of a simulated world
- Understanding criminal behavior

Simulating a realistic world

A key goal in this simulated virtual world is to create realistic data. To accomplish this goal, IBM Synthetic Data Sets leverages a broad set of statistical population data. For example, the US Census Bureau has a wealth of information down to the postal code level, with a typical address code containing 10,000 people. This information includes distributions for income, age, homeowners versus renters, monthly mortgage or rent payments, housing construction type, housing age, and other information. The US Federal Reserve supplies related information on the value and types of financial assets and debts, such as checking and savings accounts, real estate, and home, vehicle, and student loans. The Federal Reserve also presents statistics on credit and debit card spending. The US Bureau of Labor Statistics also provides a distribution of approximately 800 job types, and the pay ranges for those job types.

With this information, IBM Synthetic Data Sets builds a population whose attributes mimic the overall US population in terms of income, age, and geographic distribution. To emphasize, the simulated people that are created by IBM Synthetic Data Sets are *not* built from anonymized real individuals. Instead, the simulated people are built by using the previously mentioned statistical distributions. Although the aggregate behavior of the simulated people matches the aggregate behavior of real people, data security, privacy, or compliance risks are alleviated because no simulated individual person is based on any real individual person.

Similar to real people, every simulated person is unique. People living in the same neighborhood with similar income might have different spending habits: frugal versus expansive, high expenditures on clothes versus high expenditures on travel, and other habits. This behavior generally follows statistical patterns. For example, individuals with a higher income can afford to do more activities and have a greater tendency to spend on luxury items than someone with a lower income. However, some high-income people might spend modestly, and others spend lavishly. Low- and middle-income people also vary in their overall spending and in their specific tastes.

Creating regular and varied consumer behavior

When the simulated people and companies are created, they must participate in activities. To support these activities, IBM Synthetic Data Sets assigns other attributes, such as occupations or family size. Some of the simulated people live alone, and some are unemployed or retired. Based on their situation, people move through simulated years, months, days, and hours, and engage in different consumer behavior. For example, some people stop for coffee on weekday mornings on the way to work. The coffee purchase yields a transaction at a merchant in a specific locale. This transaction might be with a credit card, a debit card, or cash. IBM Synthetic Data Sets sees and tracks all transactions and consumer activity, which includes cash transactions. In contrast, real data often misses cash transactions. This universal data collection is one of many advantages over real data because synthetic data captures a broad, full picture of consumer behavior.

Also, IBM Synthetic Data Sets incorporates patterns and variety in consumer behavior. For example, real people's weekend consumer behavior likely differs from their weekday consumer behavior. The simulated people in IBM Synthetic Data Sets mimic this change in behavior. Simulated people take business trips and vacations at varying frequencies and spend for the destination. Simulated people spend more on gifts around certain months or holidays as well. Most simulated people are paid at regular intervals, such as weekly, biweekly, semi-monthly. Rent, mortgage, and other loan payments are typically paid once per month, with a skew toward the end of the month. IBM Synthetic Data Sets models all these details and many others with precision, which generate a realistic record of consumer behavior and spending activity.

In summary, IBM Synthetic Data Sets simulates realistic people, companies, and activity. Consumer activity and behavior follow realistic time intervals with purchases that are made on appropriate days, times, and locations.

Constructing real assets

In addition to financial transactions, IBM Synthetic Data Sets carefully models homes and other real assets. Based on census distributions, IBM Synthetic Data Sets assigns a certain home size, style of construction, and type of roofing to each simulated person. Different insurance risks are also assigned to each person and home, such as hurricanes, earthquakes, and volcanoes. These risks are based on appropriate geographical and time constraints. For example, hurricanes are more likely to hit the US state of Florida than North Dakota, and earthquakes are more likely to occur in California than in Iowa. IBM Synthetic Data Sets models the occurrence of these natural disasters with their simulated population because when a disaster occurs, home damage likely arises and leads to insurance datasets. For each claim, there is a rich set of information about exact losses, such as the home itself or loss of furniture or jewelry, and the cause of the loss, such as hurricane, fire, or theft. The claim also details exact dollar amounts in each item category and in aggregate. To enhance compatibility with databases and spreadsheets, IBM Synthetic Data Sets structures its information in tabular form and is packaged as comma-separated value (CSV) files.

IBM Synthetic Data Sets also attaches free text descriptions to each claim. This text content is generated based on exact knowledge of the underlying claim, which makes it consistent with the tabular data. For example, the tabular data might note specific items that are damaged in a flood and the loss amount for those items. The text might provide a brief description of the claim, such as “Last week my home was damaged in a flood and there is a great deal of damage to my furniture and carpets. Can you please get me reimbursed quickly for these items?”

Connecting different parts of a simulated world

Interdependence is another important aspect of how IBM Synthetic Data Sets constructs its virtual world and population. IBM Synthetic Data Sets contains a mix of over 300 large, multi-national real companies and tens of millions of small, fictitious companies. Companies can serve as both merchants that provide goods to consumers and as employers that provide salaries to simulated people. Companies can be buyers to some businesses and suppliers to others. Simulated people also contribute through consumption and investment, with their purchases increasing revenue and stock for companies. Revenue for large companies is based on the company's Form 10-K filings, and these large companies add a further element of realism to the dataset.

Understanding criminal behavior

Criminal activity is an important part of IBM Synthetic Data Sets. Having data around fraud and money laundering is imperative when training artificial intelligence (AI) models to recognize similar activity. This criminal activity includes check fraud, insurance fraud, payment card fraud, authorized push payment (APP) scams, and money laundering. The criminal activity expands to a broader set of pursuits, such as yielding illicit income through extortion, smuggling, and illegal gambling. Like other aspects of the simulated world, IBM Synthetic Data Set treats each criminal entity as unique entities with their own amounts and types of unlawful activity. Nevertheless, it is emphasized that in IBM Synthetic Data Sets only a few companies and people engage in criminal activity, that is, about 1 in 1000 or fewer.

Furthermore, with its knowledge of ground truth and universal data collection, IBM Synthetic Data Sets offers a key advantage over real data when training models to recognize criminal activity. The dataset knows who is engaged in criminal activity, when they do it, and the financial amounts that are involved. As a result, all illegal activities are identified and labeled with 100% accuracy in the dataset, which includes all scams, credit card fraud, check fraud, insurance fraud, and money laundering. With real data, this scale of illegal activity is challenging to detect. Therefore, AI models that are trained with IBM Synthetic Data Sets have a clear, accurate understanding of criminal behavior.

Legal usage terms

For the full legal terms for IBM Synthetic Data Sets, which include how to use and redistribute the datasets, see IBM Terms.

Getting started

This section describes a few different ways to get started with IBM Synthetic Data Sets:

- Artificial intelligence on IBM Z Solution Templates
- IBM Technology Expert Labs Services
- Starting a proof-of-concept with the AI on IBM Z team

Artificial intelligence on IBM Z Solution Templates

AI Solution Templates is a suite of pre-built blueprints that guide you through the full artificial intelligence (AI) lifecycle on IBM Z with various enterprise use cases while leveraging various technologies at no charge. Whether you are a senior data scientist or have no previous AI skills, you can build your own AI model, deploy it on IBM Z, and integrate it into a business application.

For more information, see AI Solution Templates on GitHub.

IBM Technology Expert Labs Services

IBM Expert Labs is a professional services organization that is powered by an experienced team of product experts. This knowledgeable team brings deep technical expertise across software and infrastructure areas. IBM Expert Labs uses proven methodologies, best practices, and patterns to help IBM Business Partners develop complex solutions and achieve better business outcomes.

There are three paid services offerings through IBM Technology Expert Labs for using IBM Synthetic Data Sets for model training and deployment:

- AI Exploration and Model Training: Integrate and blend data from IBM Synthetic Data Sets and real data, including from IBM Z and LinuxONE. Transform the data and use it for training a machine learning and deep learning model.
- Implement Machine Learning for z/OS: Install and configure Machine Learning for z/OS for model deployment on IBM Z.
- Model Deployment to IBM Z and LinuxONE: Deploy the model to IBM Z and LinuxONE for accelerated inferencing with Machine Learning for z/OS or AI Toolkit for IBM Z and LinuxONE

For more information, contact systems-expert-labs@ibm.com or your local IBM Technology Expert Labs team.

Starting a proof-of-concept with the AI on IBM Z team

Interested in getting started with a discovery workshop to discover a use case for AI on IBM Z with synthetic datasets? Want to get started on a proof-of-concept?

If so, engage with the team by reaching out to aionz@us.ibm.com.

Frequently asked questions

Here is a list of frequently asked questions (FAQ) about IBM Synthetic Data Sets:

- What are the benefits of IBM Synthetic Data Sets?

  For examples about how to leverage IBM Synthetic Data Sets for AI models and large language models (LLMs), see “Introducing IBM Synthetic Data Sets” on page 1.

- How large are the datasets?

  Each dataset comes in three editions or sizes: Trial, Pro, and Enterprise. For more information, see “Available editions” on page 7.

- What is included in the datasets?

  Information about column titles and data attributes, including examples and options, is described in “Previewing data schemas” on page 9 and can be found in the following website: https://github.com/IBM/IBM-Synthetic-Data-Sets.

- What is the methodology for creating the datasets?

  In short, the datasets are created by using the agent-based modeling method. For more information, see “Data generation methodology” on page 14.

- What environment or platforms can I download the datasets on?

  These datasets are downloadable, comma-separated value (CSV) files that are compatible with the training platform of your choice. The intention is that IBM Synthetic Data Sets can be used by IBM Z and LinuxONE customers and ISVs to build models on any platform and deploy those models back to IBM Z and LinuxONE, where the core enterprise data is for accelerated inferencing.

- How realistic are the datasets?

  IBM Synthetic Data Sets is realistic because they were created with real statistical population data from various sources, which include the US Census, Federal Reserve, Bureau of Labor Statistics, and FBI Crimes Insights, among other sources. Also, a large US national card provider compared the distribution of the datasets against their real transactions data and found that it matched well.

  Figure 1 displays the distribution of synthetic data compared to real data for payment card transactions. This data was sourced from a large US national card provider. For more information, see Synthesizing credit card transactions.

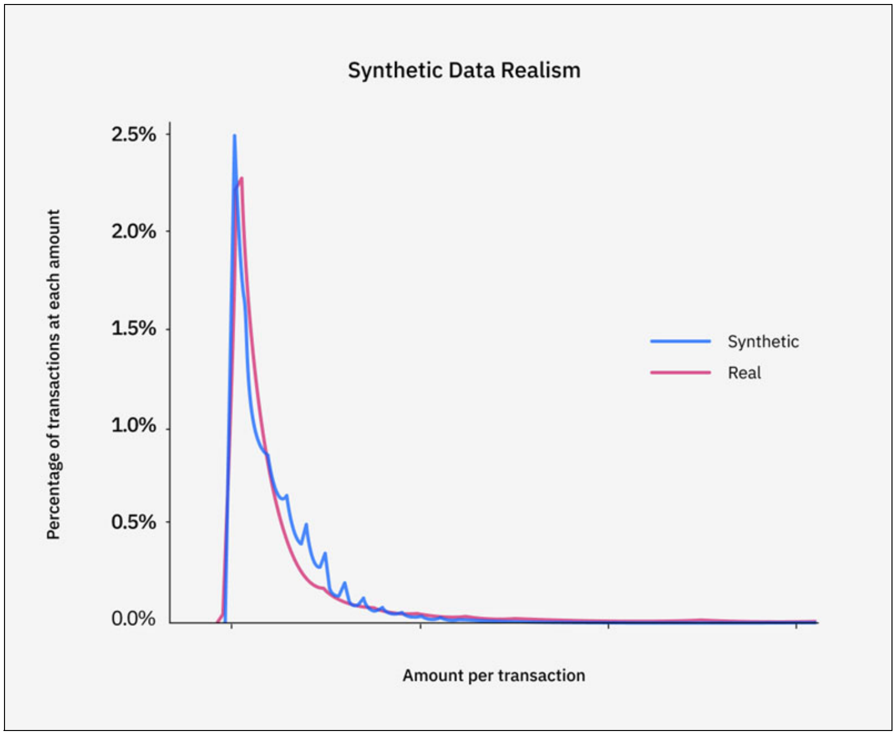


Figure 1   Synthetic Data Realism

- Will I need to transform the data?

  You might need to transform the datasets for model training or to better match the company's real data. Typical data transformation processes are permitted. For the data usage terms, read the Service description, which can be found in "Legal usage terms" on page 18.

  If you need help with transforming data, combining data sources, and training models, you can use an IBM Technology Expert Labs offering to do these tasks. To learn more about the Expert Labs offering, see "Getting started" on page 19.

- How is IBM Synthetic Data Sets different than a synthetic data generator?

  Synthetic data generators are great tools when you have access to your real data, and many of them can redact Personally Identifiable Information (PII). However, many generator tools do not produce the quality of data and logic from real data that you get from IBM Synthetic Data Sets. For example, a synthetic data generator can generate 16-digit credit card numbers but might not maintain the logic of what those numbers mean. For example, Mastercard starts with a 2 or a 5, and is aligned correctly with the column for *card company* as *Mastercard*.

  Another frequent issue with synthetic data generators is that city, state, country, and postal codes do not match in the generator outputs. For example, the city of New Orleans shows up in Italy, or Los Angeles is assigned a postal code of 2215 when only 90001 to 90042 is available. This mismatch occurs because most synthetic generators generate new data based on statistical representations from each column attribute. However, the generators do not tie into the underlying logic to produce the quality of data that is needed.

  To get the same quality of synthetic data as IBM Synthetic Data Sets, an organization would need time and money for a data scientist and a subject matter expert to spend years finding the right source data and potentially writing extra code to maintain the data logic. However, clients can promptly begin modeling and LLM training with IBM Synthetic Data Sets.

- IBM Synthetic Data Sets offers only US-based data. How does it help me if I am not in the US?

  IBM Synthetic Data Sets is most directly useful for the US. However, they can provide significant benefits worldwide:

  - The core of many AI models is pattern detection and deviations from those patterns. For example, AI models look for deviations from common or typical behavior to detect fraud and money laundering. Then, the model flags these deviations as potential fraud, or money laundering. This approach is geographically independent. If a model can find patterns in US-based data, the model is typically capable of doing so anywhere.
  - The patterns are geographically independent. For example, it is always unusual to have multiple purchases in an hour at brick-and-mortar merchants when the merchants are separated by hundreds of kilometers. It is always unusual for someone who spends frugally to suddenly spend large amounts on expensive luxury items. Certain patterns of transfers between bank accounts are common, such as moving money from checking to savings. Other patterns might be less common, such as suddenly moving small amounts of money to a large set of other accounts. As a result, although IBM Synthetic Data Sets is US-based, the logic behind pattern detection and deviation can be applied universally.

    Patterns might be more subtle than these examples. Use broad, well-labeled data to create and train AI models to detect such subtleties.
  - The data generation that is used for IBM Synthetic Data Sets simulates international companies and business transactions worldwide. The simulated people and companies travel and conduct transactions in 223 countries around the simulated world, and use international currencies and banks to facilitate their activities. Therefore, although the datasets' transactions center is in the US, they cover the world.

  IBM Synthetic Data Sets has many attributes that are not available in real data. IBM Synthetic Data Sets has fully accurate labeling for a broad set of categories. IBM Synthetic Data Sets also provides data for all banks and insurance companies in the ecosystem, which includes cash transactions that are frequently overlooked by real data.

  Clients can combine IBM Synthetic Data Sets with local data to develop enhanced, robust capabilities that are beyond what IBM Synthetic Data Sets or local data alone can independently offer. IBM Synthetic Data Sets can also fine-tune models that are created from local data.

- If I have feedback on how to improve the datasets, how do I provide that feedback?

  We appreciate your feedback and aim to include relevant suggestions in future updates to the datasets. Updates are available with the purchase of a subscription service.

  To submit new ideas, see ideas.ibm.com.

Additional resources

- For more information about synthetic datasets, see the following resources:2021 International Conference on AI in Finance (ICAIF): Synthesizing credit card transactions
- 2024 ICAIF:
  - FraudGT: A Simple, Effective, and Efficient Graph Transformer for Financial Fraud Detection
  - Graph Feature Preprocessor: Real-time Subgraph-based Feature Extraction for Financial Crime Detection
- 2023 Neural Information Processing Systems (Neurips) paper: Realistic Synthetic Financial Transactions for Anti-Money Laundering Models
- 2024 Association for the Advancement of Artificial Intelligence (AAAI) paper: Provably Powerful Graph Neural Networks for Directed Multigraphs

REDP-5748-00

ISBN 0738461997

Printed in U.S.A.