Deep Science Publishing, 2025
https://doi.org/10.70593/978-93-7185-061-2_5

DeepScience
Open Access Books

Chapter 5: Data Engineering for Risk and Fraud Intelligence

5.1. Introduction to Data Engineering

Data engineering involves designing and building scalable infrastructure, architecture, and workflows for ingesting, processing, and storing very large volumes of data. It is a prerequisite for data science, machine learning, and big data analytics. Data Engineering for Risk and Fraud Intelligence applies these general principles to the specifics of risk and fraud intelligence. Investigations into specific data sourcing techniques, storage options, and risk analysis methods help generate high-quality data to feed risk models. Many of the risks that organizations face are related to the risk of fraud and financial crime. Fraud consumes an inordinate amount of an organization's resources and exposes the firm to considerable losses and regulatory penalties. These are some of the primary objectives of risk and fraud intelligence: detecting money laundering; mitigating fraud losses; understanding the identity of the persons you are investigating; understanding the activity undertaken by those persons; and determining the risk associated with those persons.

Risk and fraud intelligence require a vast amount of different data types—structured and unstructured. Data can be utilized straight from a database, or can be extracted by scraping websites, using external data feeds, or tapping into social media URLs. These data streams can take the form of either batch or real-time, depending on the forensic and analytics requirements. Storage options range from online, transactional systems to data lakes or federated data warehouses. Increasingly, storage is moving to the Cloud for considerations regarding security, scale, location, and regulatory compliance. Risk and fraud models can be generated through batch or stream processing paradigms. Data pipelines can be built using either an Extract, Transform, Load (ETL) or Extract, Load, Transform (ELT) approach. The following sections explore these aspects in more detail. (For an overview of data engineering, see Section 1.1 Data Engineering Fundamentals.)

5.1.1. Overview of Data Engineering Fundamentals

Data engineering focuses on the practical application of data collection, processing, and storage. It concentrates on the infrastructure and related workflows required to handle data coming in from various sources in large volumes and at velocity, often defined as big data. Several different data sources can feed the risk and fraud intelligence models. Web scraping techniques are used to collect data from targeted websites, search engines, and RSS feeds, and different sources such as APIs and databases. A Data Lake can be used for storing the data collected in the raw format. A Data Warehouse can be used for storing the transformed data and the results from the risk and fraud intelligence models [1-3]. Batch and stream processing concepts are introduced and their differences highlighted in order to be able to propose a possible architecture. The differences between ETL and ELT are also highlighted.

The primary objective of risk and fraud intelligence is to detect fraudulent behaviour and attacks that can result in financial loss or identity theft. Many sectors face the risk of fraud, but in particular banking and payments where the risk of financial loss can be very high. Different data sources can be used to feed the models to help detect these types of attacks. The data sources can be a variety of structured, semi-structured, and unstructured data.

5.2. Understanding Risk and Fraud Intelligence

The primary objective of risk and fraud intelligence is to detect and analyze inconsistent behaviors and anomalies that indicate fraudulent or malicious activities. This detection may target financial frauds and related threats, such as money laundering and bribery, or aim to identify security attacks, including network overload and cybercrime. Organizations rely on risk and fraud intelligence to mitigate vulnerabilities and prevent significant damage to operational, reputational, or market valuation aspects of their business.

These detection and analysis goals require extracting signals from a broad range of data sources, acknowledging that fraudsters craft schemes using varied information sources. Consequently, the data necessary for risk and fraud models extend well beyond the organization's internal data.

5.2.1. Key Concepts in Risk and Fraud Intelligence

Understanding Risk and Fraud Intelligence focuses on the goals of risk and fraud intelligence and examines the data engineering aspects that underpin these objectives. The chapter defines the term and proceeds to identify and describe typical data sources.

Methods for gathering risk data are then introduced, followed by a discussion of storage solutions for accumulating risk data. The final topics address processing requirements, processing methods, and future challenges in the area.

Risk and fraud intelligence share the overarching goal of identifying individuals or activities that could pose a risk or threat. Although the objectives are similar, their focal points differ: fraud intelligence centers on determining whether specific individuals are carrying out fraudulent activities, while fraud detection aims to recognize particular activities as fraudulent. Security intelligence involves identifying individuals who present a risk for malign activity and assessing the severity of that risk. Anomaly detection is the broader concept of identifying unusual activities in specific domains, which can then be used to calculate risk scores. Risk and fraud intelligence and other forms of anomaly detection require the analysis of multiple samples, each encompassingn attributes or features, with the objective of categorizing each sample as low risk or high risk.

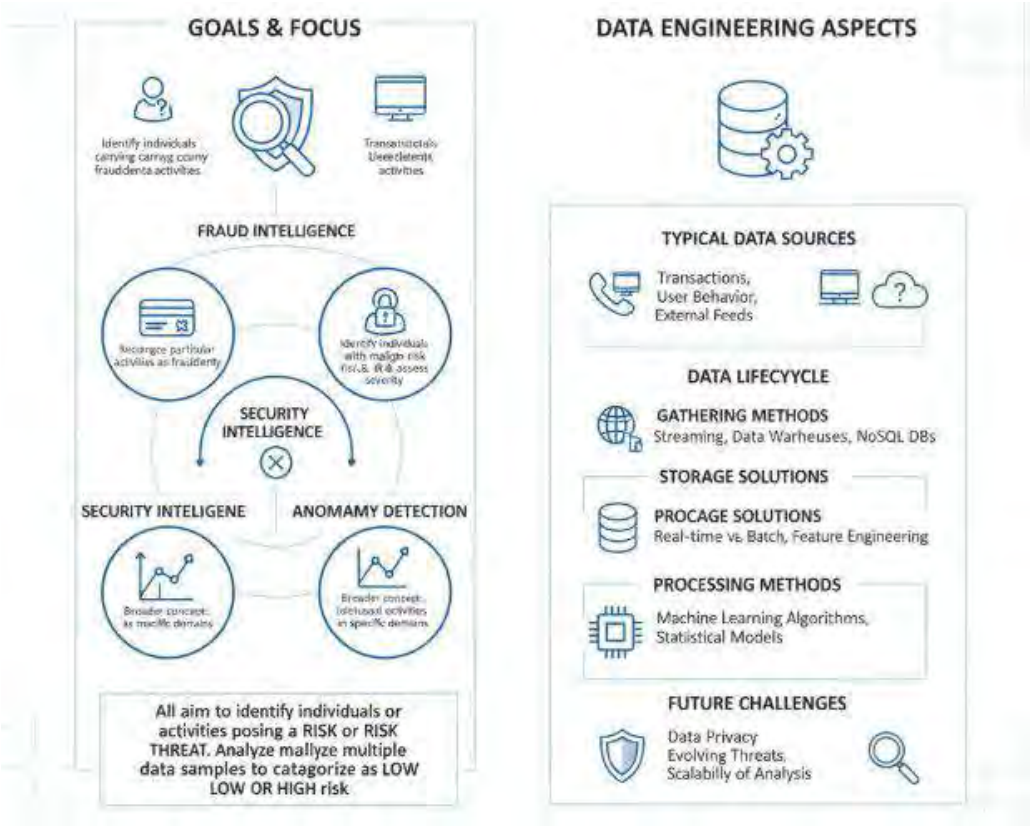

Fig 5 . 1 : Goals and Concepts of Risk and Fraud Intelligence

5.3. Data Sources for Risk Analysis

Risk and fraud intelligence glean insights into situations and processes that carry a higher probability of adverse consequences, enabling mitigation and planning strategies. Drawing parallels with cybersecurity, risk and fraud share a focus on protecting assets through real-time assessment, detection, and preemptive action against potentially harmful events such as theft, fraud, and money laundering. Although numerous public datasets are available, few specifically support fraud or risk-related endeavors. Data engineering integrates established principles with the practical requirements necessary to fashion effective risk and fraud intelligence systems.

The various data sources utilized in fraud and risk intelligence projects include internal databases acting as digital breadcrumbs, social media data providing behavioral clues, Government Open Data offering validated, extensive datasets, subscription databases concentrating on geographic or industry-specific risk factors, incident and breach databases recording conflict and crime data, and dark Web data supplying contextual inferences. Fraud information also emerges from the unstructured content of news reports. New entrants capable of providing income risk assessments have identified an opportunity to capitalize on the creditworthiness of the "near-prime" consumer: the estimated 150 million consumers in the USA and 300 million in the EU who are above the "deep sub-prime" and "sub-prime" risk bands, and yet do not meet the scratch scoring models of the banks.

5.3.1. Internal Data Sources

Risk and fraud models for money laundering, credit risk, or anti-fraud systems are fed by vast quantities of digital data. This data can take a variety of forms, from structured databases to unstructured media such as photographs or sound. Consequently, the engineering challenge lies in building data pipelines that provide the right types of data to risk models in the right place, at the right time, and in the right order of magnitude. These data sources are not processed for the benefit of the risk models but for a variety of uses demanded by the bank, for instance, regulatory reporting, transaction reporting, trade finance, and many others. Financial institutions base their operational systems on historical data required for their day-to-day operations and also use their data for the purpose of gaining insights through machine learning and business intelligence.

Data is captured within the organisation and sourced from external business partners, including customers and intermediaries. Internal data such as customer data, account data, watch lists, and other details are essential for processing. There is a wide variety of operational systems for storing data, ranging from simple legacy databases to FTP and batch systems. However, the methods of data sourcing for risk and fraud models also

vary widely. Data engineers for risk and fraud rely heavily on sourcing the data through Structured Query Language (SQL) tables as internal data feeds, as well as using web scraping techniques and Application Programming Interfaces (APIs) to create these feeds. Through open protocols, many companies have exposed their transactional data in real time to the outside world. These are referred to as real-time data streams. The ingestion of real-time data streams into the bank and their combination with web-scraped data help the risk models in detecting fraudulent behaviour [2,4,5]. Once data has been collected, it can be stored in a data lake or a data warehouse. The use of cloud storage through cloud vendors such as Amazon Web Services can be an efficient method with multiple in-built features for risk and fraud.

5.3.2. External Data Sources

The models necessary for a comprehensive understanding of fraud and risk seldom reside on internal systems alone. External data sources augment models with the external characteristics of risk entities. Surrounding the customer profile with non-company-

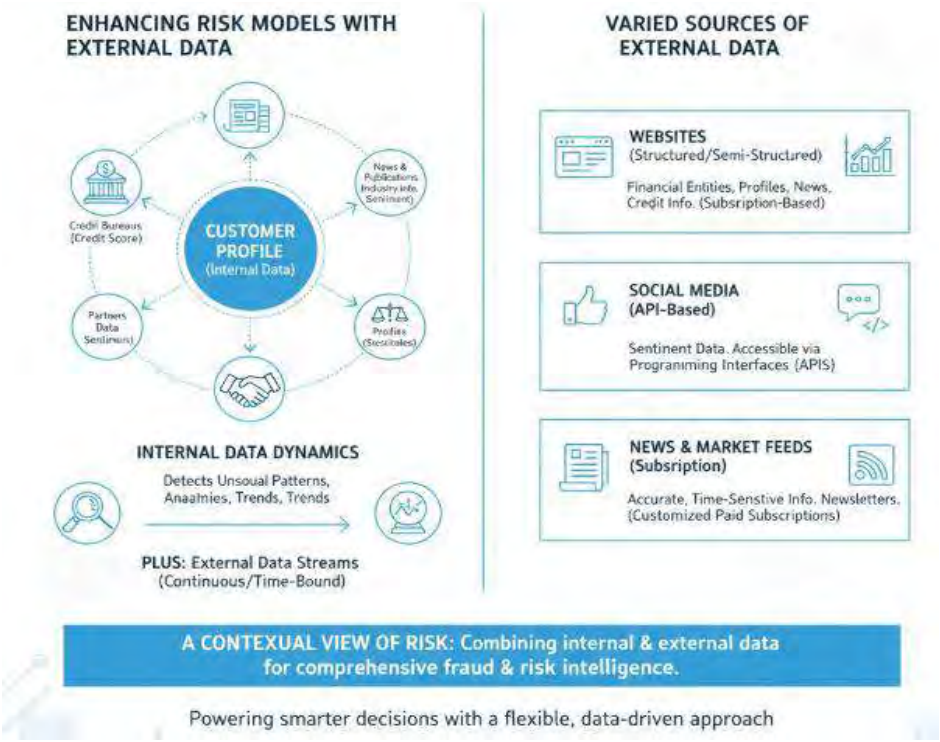

Fig 5 . 2 : External Data Sources for Risk and Fraud Intelligence

generated information—such as credit score from a credit bureau, news published about the customer or their industry, information about their partners and associates, social media profiles, and so on—enables models to provide a contextual view of risk. Internal data is not necessarily static, either; fraud and risk intelligence systems aim to detect unusual or abnormal patterns, anomalies, trends, or even predict future behaviors. External data streams, when captured continuously or within specific time bounds, also offer an important dimension.

The sources of external data for risk models are highly varied. Typically, websites offer structured or semi-structured data about financial entities, ranging from profiles and news to complex metrics such as credit scores and detailed legal information. This information, though curated and presented by a third party, is usually available as part of a subscription model. Other sources may include social media sentiment data, rarely accessible through direct database connections but instead made available via application programming interfaces (APIs). Lastly, several financial market and business news providers, who maintain a reputation for delivering accurate, time-sensitive information, offer newsletters and feeds. Customized, paid-for subscriptions to these data-streams can be included as part of risk workflows.

5.3.3. Real-time Data Streams

Data relevant to risk and fraud analysis originates from a wide range of sources and in many forms. It is not just structured data from databases, or unstructured data such as text and images—data that can be probing and sometimes personally intrusive—but also processes, relationships, behaviour, and actions. Insights can be found in what people do, not only in what they say or in what they have. The variety of data is a major challenge, as is the velocity at which it can be collected and delivered. Fraud detection has to be as close to real time as possible as many threats are executed in seconds—only a short window is available to identify and block action before it occurs.

Extracting structured data from websites allows frequent scraping and comparison over time. Application programming interfaces (APIs) provide a real-time feed from a target site or source, as do RSS feeds and data streaming [1,3,4]. Databases are accessed to extract information, typically on a less frequent basis, but still critical as the data source. Some data providers produce a data file for upload and use in a system or application. To take advantage of all such sources, storage solutions need to be flexible and scalable. Data lakes offer a solution, as raw data from a variety of input sources can be stored and used later for processing and delivery into staging tables in a risk data warehouse.

5.4. Data Collection Techniques

The data sources feeding risk models are numerous and diverse. Internal and external organizations operating across multiple geographies capture and process vast amounts of fraud-related data in transaction logs, customer profiles, social media, online websites, and in telecommunications networks. These data bytes represent accumulative knowledge of entities engaged in multiple forms of transactions. Natural events such as droughts, excessive rains, or earthquakes that have an impact on risk to human lives also provide important information stored in dedicated repositories that are utilized for risk analysis.

Analytics teams are interested in collecting data in the form of flat files or in-memory data structures at periodic intervals from various sources through different pipelines. From the data engineer's perspective, the data collection methods must be able to set up pipelines to routinely ingest the discovered datasets. Web scraping, Application Programming Interfaces (API), Data Feeds, and Database Extraction are common methods for reliably collecting data originating from external sources.

5.4.1. Web Scraping

A broad risk model relies on collecting data and information relevant for identifying, investigating, mitigating, and reporting risk. The choice of sources can span a broad spectrum, including web scraping services, APIs and data feeds, or database storage.

Web data extraction involves capturing and transforming selected data from various websites into organized formats for further refinement and analysis. The necessity for web scraping stems from the persistent growth of information available on the internet. Businesses and organizations require real-time access to such information for diverse applications. The dynamic nature of the World Wide Web, characterized by the adjustment of web page contents, ensures that web scraping meets the evolving needs of various enterprises. The retrieval of extracted content is supported by tools such as Scrapy—a Python framework designed for efficient and scalable web crawling and scraping—and Beautiful Soup, a Python library specialized in parsing HTML and XML documents, facilitating convenient data extraction.

5.4.2. APIs and Data Feeds

Extracting data from structured sources is often straightforward. Public data in HTML tables or lists can be extracted using specialized web scraping techniques. Many banking and fraud risk control applications rely on external data from social media, credit bureaus, or data aggregators. They often provide Application Programming Interfaces

(APIs) or subscription data feeds that enable real-time risk evaluation. Internal corporate data sources, such as data warehouses or Enterprise Resource Planning (ERP) systems, offer rich insights and can be conveniently accessed via APIs or through Extract, Transform, Load (ETL) pipelines.

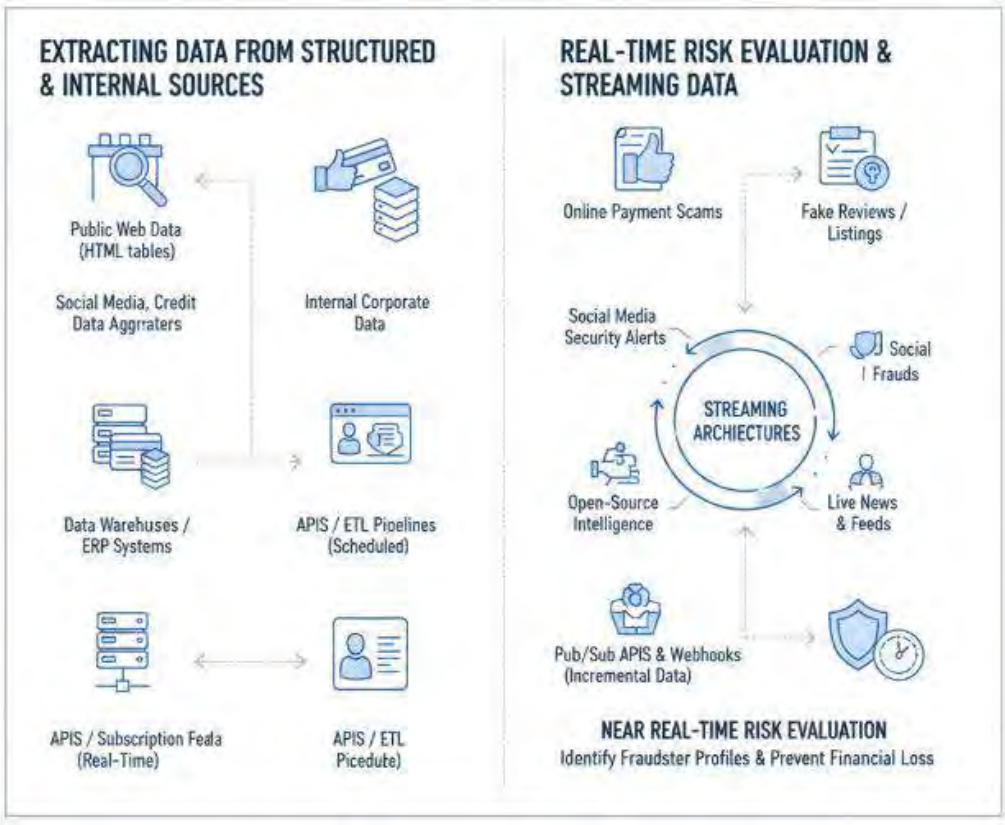

Fig 5 . 3 : Data Extraction Methods for Risk and Fraud Intelligence

The growing variety of fraud techniques, fraudster digitization, and the emergence of new fraud schemes such as online payment scams, fake reviews, scam listings, fake product sales, and social frauds have driven the requirement for near real-time risk evaluation. Streaming architectures based on event data—including social media security alerts, open-source intelligence, live news and social media feeds—present a new generation of risk inputs that can be used alongside traditional batch-based models to identify fraudster profiles and prevent financial loss. These sources are typically accessed via pub/sub APIs and webhooks, delivering data incrementally over time.

5.4.3. Database Extraction

Data used for risk or fraud intelligence can also be collected from internal databases. This category of data includes recipes, recipes_ingredients and recipes_nutrition, whereas capturing structured data manually or automatically via APIs is more complicated.

Many companies manage these types of information using internal databases, which are then accessed via third-party platforms. These centralized databases, or specific views of centralized data collections, are available for queries and can provide data of higher quality with more reliable access times; generally, they are used without the need of moving the data. Nevertheless, different internal systems support these platforms. In data engineering, data lakes hold raw data in its original format, whereas data warehouses employ a highly structured architecture.

5.5. Data Storage Solutions

Data feeds for risk and fraud analysis come in large quantities and unbounded velocities and in multiple formats [6-8]. Data storage solutions used in such scenarios need to provide efficient querying of analytical data and flexibility of ingestion with little or no concern for later use. Such a system needs to be easily scalable so that growing data volume can be stored reliably, efficiently and securely.

Risk and fraud analysis relies on both structured and unstructured data, which in turn require different storage architectures. Data lakes emerged as one solution with the rise of big-data, machine-learning, artificial intelligence, and chatbot tools, which consume more unstructured data. A data-lake stores data in its raw form, providing flexibility to query at a later stage. In contrast, a data-warehouse stores data in a highly structured form to support large volumes of simple queries efficiently. Specifically, data-warehouses are a specialized version of relational database management system (RDBMS) designed to hold large amounts of current or historical transactional data for analysis. A wide range of service providers offer both data-lake and data-warehouse architectures in the cloud. Cloud platforms provide low cost, high-performance options for data storage, with powerful kernels to handle data privacy and industry regulations.

5.5.1. Data Lakes

When developing data engineering processes for risk and fraud intelligence, it is important to choose data stores that can handle the enormous volumes of information generated by source systems. Data lakes provide a cost-effective solution for storing very large datasets within their native structure.

The aforementioned collection techniques result in a myriad of distinct datasets. However, because they share a common purpose of supporting risk intelligence, they can all be combined into a dedicated data store. At first, this mixture of organisations, register information, transactions and events will be in no more than a giant data soup, so working with it may be quite challenging. But it gets easier and more useful as it is gradually organised and combined into analytical datasets.

5.5.2. Data Warehouses

A data warehouse is a database designed to store vast amount of structured data for analysis and reporting. The primary difference between data lakes and data warehouses is that data lakes store unstructured, raw data, while data warehouses serve as repositories for highly structured data and are optimized for analysis and querying. Despite the differences, both lakes and warehouses play crucial roles in the data processing chain for risk and fraud intelligence.

Data lakes are usually cheaper than data warehouses because they store raw data on cheap object storage services. They also enable a wider range of operations since raw data can be used for all types of analysis, including discovering unknown patterns. Data warehouses, on the other hand, are utilized for fast personalized or massive data extractions and reporting, as they use a highly structured format that is easy to analyze with SQL. Storing highly structured data in a data lake is expensive, and storing unstructured data in a data warehouse fails to take advantage of the main strength of the warehouse.

5.5.3. Cloud Storage Options

Security, scalability, durability, regulatory compliance, and cost are critical requirements when storing risk analytics data. Cloud computing offers infrastructure more suitable because the cloud allows for rapid scaling, reducing the risk of disruptions during data load. Elastically scaling infrastructure both vertically and horizontally is a tremendous advantage. Vertical scaling (changing the size of a virtual machine by adding or removing cores, disk, or RAM) allows accommodating short bursts of workload-time demand. Horizontal scaling (adding or removing virtual machines) is necessary for long-term workload demand increase or decrease. Using the cloud, it is also possible to leverage additional durability of the dataset by having the data replicated in multiple locations within a geographic region or across geographic regions.

Cloud services make it possible for each organization to have a good-enough disaster recovery process, with a Recovery Point Objective (RPO) of minutes rather than hours

or days — even though in an on-premises setup that might be inconceivable for many businesses because of the near-doubled infrastructure cost. Besides scalability and durability, the cloud also makes it possible to create an event-driven architecture that can optimize processing cost. It is essential to mention that, depending on the sensitivity of the data, organizations might still decide to keep the dataset on-premises or hosted within a cloud service. The final decision depends on the business's centricity toward the cloud and the regulatory requirements.

5.6. Data Processing Frameworks

The choice of a processing framework for transforming raw data into usable information depends on business intelligence needs. Establishing workflows is usually necessary. Risk and fraud intelligence analysis is often performed on data that has been extracted, transformed, and loaded. Changes in data content can trigger the transformation process. Although databases normally process extract–transform–load operations, certain risks demand different treatment of the framework, adopting an "extract–load–transform (ELT)" methodology.

In fraud detection, the increase in streaming data fuels the growth of stream processing. Risk and fraud intelligence involve Low Latency and Very Low Latency analyses for swift anomaly detection and prompt reduction of risk exposure. Hence, risk and fraud intelligence systems frequently adopt stream processing frameworks.

5.6.1. Batch Processing

Intuitively, batch processing involves ingesting large volumes of data, mutually transforming them, and subsequently storing the results for downstream querying and visualization. The entire batch usually needs re-processing whenever new records arrive. Such an approach tends to be preferred whenever applications are not sensitive to the data processing latency; in fact, some advantage can be derived from higher latency, such as flexibility to incorporate a wide range of data cleansing, transformation, and enrichment techniques without drastically increasing the cost and complexity of the data engineering pipelines.

Scenarios where input data is susceptible to sudden changes in velocity, volume, and variety also overall benefit from batch processing systems since they can be designed to scale out without affecting the latency but possibly increasing the execution time of the batch processing jobs. The time-bounded nature of batch processing makes it possible to plan execution so that it does not interfere with other business-critical processes, such as

online customer-facing services. For many organizations, such a practical constraint might alone be the reason for batch-oriented architecture.

5.6.2. Stream Processing

Data analytics for risk and fraud intelligences typically exists on a continuum from full real-time to batch mode and back again as the risk/detection team evolves. The latter, batch mode, is naturally simpler and requires less network, cloud, service, and cloud-computing resource expenditure, although it will miss strategic payments or orders and operational business damage. When operated in a streaming manner, it can identify an emerging incident and capture it in near real-time, offering operational protection to the business at that moment. Streaming analytics either involve processing real-time transaction data or real-time big-data web-scraped information such as website outages for a predictive operational risk model.

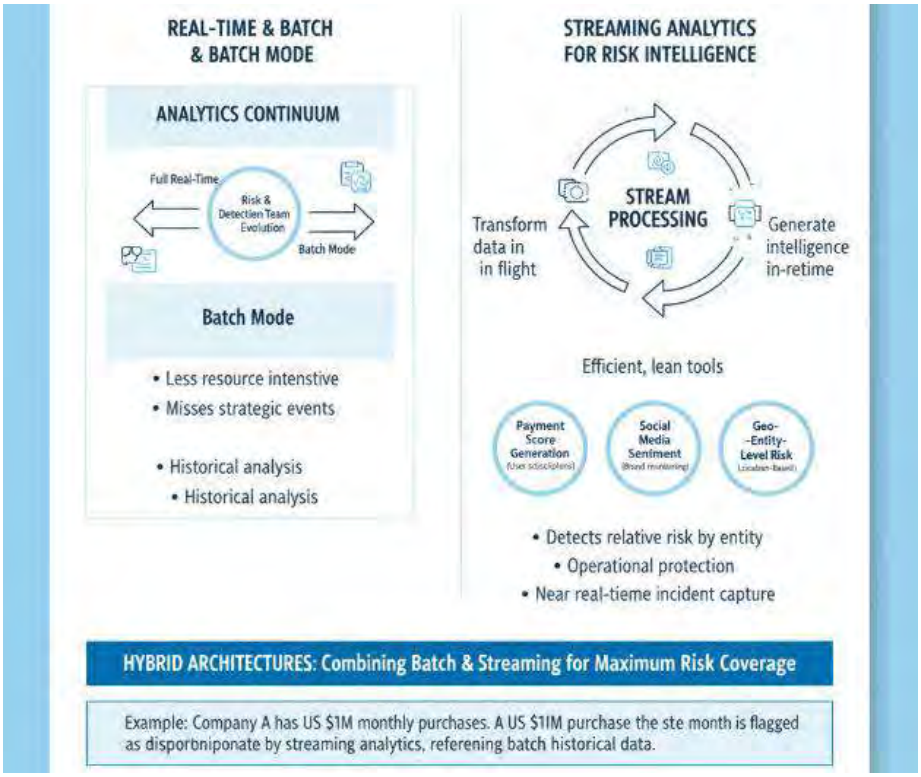

Fig 5 . 4 : Batch vs. Streaming Analytics for Risk and Fraud Intelligence

Stream-processing models transform data whilst in flight so that intelligence can be generated in real-time even before data has been stored. This choreography creates very efficient, lean tools for risk and fraud intelligence. Stream processing is, therefore,

suitable for a wide range of use cases, such as model-score generation of payments to compare user subscription levels for typical transactions, social-media sentiment detection on company brand, or geo-entity-level risk detection. While streaming analytics are expressed here as separate arquitectures, risk and fraud analytics models such as these can often combine batch and streaming data sources for maximum use of information and risk coverage. For example, some types of data will only reside in batch form, such as quarterly sales data, while credit-card transactions need to be scored at the time of authorization. For example, Company A might make purchases in excess of US $1 million per month; therefore, when a US $1 million purchase request comes through at the start of the month, the alerting system should detect that the request is disproportionate in comparison to historical purchases and flag this out automatically as a potential alert. Streaming analytics thus detect the relative risk by entity.

5.6.3. ETL vs ELT

Traditional batch processing collects data from different sources, moves it to a centralized location, and prepares it for analysis. This process may be scheduled periodically at the end of the work day and result in a day's lag between data generation and analysis. This method is known as Extract, Transform and Load (ETL), in which data is extracted from source systems, transformed to a certain format, then loaded into a target system. In the days before widespread use of cloud computing, loading data into an already transformed state was a natural choice because the processing and storage capacity were typically built into the same on-premise system.

As cloud computing became widespread, processing and storage tasks could be split up and distributed across different systems in different locations, connected through a high-speed network [7-9]. The use of an ELT (Extract, Load and Transform) approach emerged, in which data is first pulled out of source systems, then loaded into the analysis environment where the requisite transformations occur. ELT supports greater flexibility, enabling organizations to keep their data in a raw state and extract new information from it as needed.

5.7. Future Trends in Risk and Fraud Intelligence

The rapid evolution of risk intelligence technologies is shaping the future of risk management. Emerging areas such as explainable artificial intelligence (XAI), Blockchain, and smart contracts have the potential to significantly alter risk and fraud intelligence landscapes. Current systems struggle with detecting and preventing application-layer fraud, increasingly prevalent in financial services. Advanced machine learning and artificial intelligence models can substantially improve the detection and

prevention of fraud typologies, including transaction fraud, account takeover, triangulation fraud, and identity fraud.

Risk intelligence models depend on extensive data spanning various domains, necessitating tailored data engineering workflows. Innovations in data engineering tailored to the specific requirements of risk intelligence can make a considerable difference. These future directions are examined from both data and advanced analytics perspectives. From a data standpoint, aspects such as live and near-time data sourcing and storage address the requirements of emerging use cases and advanced analytics use cases discussed later. Live data sourcing enables real-time creation of associated risk models.

5.7.1. AI and Machine Learning Innovations

Following a broader consideration of emerging technologies, a closer examination of AI-enhanced data engineering reveals improvements in data collection, pattern detection, and alert generation. Data engineering must account for the increasing prevalence of artificial intelligence in fraud and risk intelligence. Large language models underpinning publicly accessible chatbots—such as ChatGPT, Bard, Bing Chat, and others—can be employed by criminals to craft more effective phishing campaigns, deceptions, and other attacks, thereby increasing the frequency and complexity of attacks.

Conversely, AI offers practical tools for risk and fraud intelligence. Chatbots can generate code snippets for web scraping or template code for database access, deduplication, and more. Semantic search can be used to index news and alerts and respond to natural language questions posed by risk analysts. Other generative AI models can create synthetic data for algorithm testing, training, and sharing without violating data regulations. These AI tools enable risk and fraud intelligence teams to work more efficiently and with greater impact.

5.7.2. Blockchain Applications

Disruptive risk and fraud intelligence innovations include widespread adoption of new technologies. Within the financial sector, the interest in blockchain technology stems from its potential to improve efficiency by removing intermediaries, modernize business models to reduce costs and increase revenues, and enhance compliance controls by providing secure and immutable logs [4-6,10]. The transparent and auditable nature of blockchain provides promising opportunities for risk and fraud intelligence, as it simplifies compliance audits, detects fraud or money laundering activities, and allows

regulators to monitor sensitive payment and transaction data in nearly real-time. The costs of financial crimes in the UK are staggering (over £190 billion in 2020 according to reports by UK Finance) and are a primary driver behind the search for emerging risk and fraud intelligence technologies.

The property market is an example of a sector that could benefit from the introduction of blockchain solutions, where large market events can have a severe impact on the wider economy due to the vast sums involved. Within this sector, there are recognized sources of fraud risk such as title fraud, money laundering, and mortgage fraud, which have the potential to be reduced by providing an immutable record of ownership and a tamper-proof audit trail, thereby assisting risk and fraud intelligence efforts in overcoming the sources that allow such crimes to exist. Moreover, regulatory bodies recognize the advantages of using distributed ledger technology as a potential solution for assessing the creditworthiness of prospective clients in a privacy-preserving manner.

5.7.3. Regulatory Changes

Expressed concerns about discrepancies between industry practice and regulation are evident in the areas of data protection, process or model risk, algorithmic risk, explainability support, data and model audit trails and provenance. Regulatory evolutions increase requirements for risk and fraud analysis systems, which consequently must also adapt their data engineering layers accordingly.

Typically, the biggest impact lies with late-stage risks such as compliance, model risk, and game-theory-based risk, adaptive fraud prevention, and anti-money laundering. These often require regulations for the sophisticated use of data, appraisal of model building and deployment tools and techniques, description of ongoing model risk monitoring activities, review processes, stress testing or scenario analysis, safeguards against gaming, well crafted explanations, and a complete audit trail of internal and external data, model decisions and actions.

5.8. Conclusion

The dynamic nature of risk and fraud intelligence demands continuous enhancement in data engineering. Embracing artificial intelligence and machine learning promises automation of complex tasks and increased accuracy in threat detection. Blockchain technology offers immutable records and transparent transactions, providing inherent fraud resistance, yet also enables new nefarious activities that countermeasures must address. Furthermore, evolving regulatory requirements, exemplified by frameworks

such as GDPR, propel developments in data engineering towards greater transparency and data sovereignty.

Understanding the interplay between various data engineering techniques and the practice of handling risk and fraud intelligence lays a foundation for navigating these future trends. Adequate preparation equips the next generation of researchers and practitioners to harness emerging opportunities and confront forthcoming challenges in these critical domains.

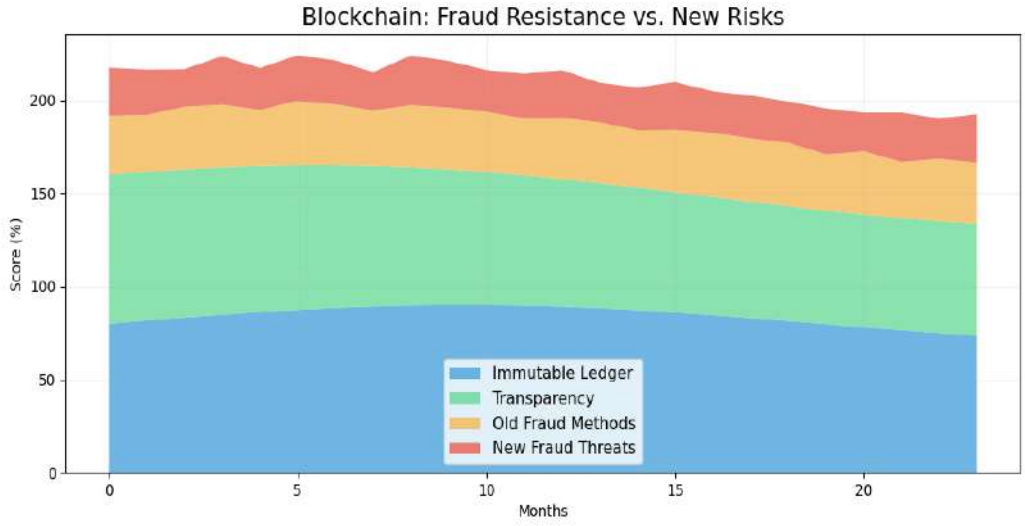

Fig 5 . 5 : Blockchain: Fraud Resistance vs. New Risks

5.8.1. Summary and Key Takeaways

This chapter demonstrated how data engineering techniques used for creating high-value data-processing pipelines can be adapted to the risk- and fraud-intelligence domain in order to ingest, handle, and transform data at scale. As fraud is a very volatile problem with regard to time and place, requiring constant action from financial institutions, the next-generation systems that are replacing nowadays legacy applications require state-of-the-art data-engineering architectures and concepts to continue protecting businesses and their customers effectively.

The study revealed how risk and fraud intelligence can be defined by listing key concepts, the main types of data, necessary risk-intelligence data sources as reported in the literature, the most common methods used to collect data, and how the generated information should be stored and processed in order to optimize the architecture of a risk-intelligence system. An overview of next-generation risk intelligence was provided by listing applications of future technologies that could potentially influence this field,

such as artificial intelligence, machine learning, regulations, technologies for securing data transmission, and the use of blockchain.

References

[1] Lee J, Bagheri B, Kao H-A. (2015). A Cyber-Physical Systems architecture for Industry 4.0-based manufacturing systems. Manufacturing Letters, 3:18–23.

[2] Sheelam, G. K., Meda, R., Pamisetty, A., Nuka, S. T., & Sriram, H. K. (2025). Semantic Negotiation Among Autonomous AI Agents: Enabling Real-Time Decision Markets for Big Data-Driven Financial Ecosystems. Metallurgical and Materials Engineering, 31(4), 587-598.

[3] Wollschlaeger M, Sauter T, Jasperneite J. (2017). The Future of Industrial Communication: Automation Networks in the Era of Industry 4.0. IEEE Industrial Electronics Magazine, 11(1):17–27.

[4] Meda, R. (2025). Optimizing Quota Planning and Territory Management through Predictive Analytics: Segmenting Sales Reps and Accounts within National Sales Zones. Advances in Consumer Research, 2(4), 443-460.

[5] Rüßmann M, Lorenz M, Gerbert P, Waldner M, Justus J, Engel P, Harnisch M. (2015). Industry 4.0: The Future of Productivity and Growth in Manufacturing Industries. Boston Consulting Group, 2015 Report.

[6] Inala, R. (2025). A Unified Framework for Agentic AI and Data Products: Enhancing Cloud, Big Data, and Machine Learning in Supply Chain, Insurance, Retail, and Manufacturing. EKSPLORIUM-BULETIN PUSAT TEKNOLOGI BAHAN GALIAN NUKLIR, 46(1), 1614-1628.

[7] Qin J, Liu Y, Grosvenor R. (2016). Digital Twin and Cyber-Physical Systems in Industry 4.0: A comparison. Journal of Manufacturing Science and Engineering, 138(4):041018.

[8] Kalisetty, S. Leveraging Cloud Computing and Big Data Analytics for Resilient Supply Chain Optimization in Retail and Manufacturing: A Framework for Disruption Management.

[9] Jeschke S, Brecher C, Meisen T, Özdemir D, Eschert T. (2017). Industry 4.0: Challenges and Solutions for the Digital Transformation and Use of Smart Automation in Manufacturing. Springer International Publishing.

[10] Revolutionizing Automotive Manufacturing with AI-Driven Data Engineering: Enhancing Production Efficiency through Advanced Data Analytics and Cloud Integration . (2025). MSW Management Journal, 34(2), 900-923.