Deep Science Publishing
https://doi.org/10.70593/978-81-988918-1-5_9

DeepScience
Open Access Books

Chapter 9: Enhancing fintech products with data engineering and machine learning pipelines

9.1. Introduction

The term financial technology, or fintech for short, is an umbrella definition that is becoming progressively popular in describing varieties of technology-driven innovation and problem-solving new business models within the financial services industry. Internationally dollars invested in fintech companies have increased markedly in recent years, and this focus on bank-assisting technologies has translated into the rise of a new and sophisticated breed of investment bank, termed the fintech bank, serving the needs of the hedge fund and private equity providers of risk capital who play a critical role in the creation of many of these newly-minted fintech companies (Chen et al., 2014; Breck et al., 2017; Akhgar & Sutan, 2021). The technologies available for use by industry new entrants is incredibly diverse—ranging from robotics, to artificial intelligence and machine learning, to blockchain and electronic market making, cloud computing, and applied neuroscience—and the fintech area is distinguished in that it is especially open to new firms and new partnerships, creating the potential for a new alliance-based buty area, particularly fevered discussion appearances by central bankers, growth and exit strategies of leading fintech firms, consumer interest change to create new forms of joint sector partnership models, fintech services as the latest product extension, and thus growth area, for the traditional bank, as well as a major accelerator of financial inclusion in developing markets (Kelleher et al., 2015; Gai et al., 2018).

9.2. The Role of Data Engineering in Fintech

Data Engineering plays an integral part in facilitating the huge growth we have seen in Fintech with the rise of multiple new ventures and innovative products. A data engineer will build the machinery that will allow data scientists and machine learning engineers to be able to accelerate their work and let them focus on solving the problems while removing roadblocks related to how we can easily run experiments at larger or production scale or how do we use large amounts of data efficiently both for AI and business analytics purposes. This has two main aspects: Engineering the machine Learning pipelines: These pipelines are how do we take the raw data coming into the system and transform it into structured datasets, how do we do it efficiently and repeatedly for both training ML models and using the models in production, and then how do we scale these pipelines up as more data comes. Supporting the data discovery and usage for business use cases which do not involve AI models: This would involve building the data warehouses where we can query and combine data easily, designing data models and ultimately dashboards to find insights into what is happening with our business which are crucial for any startup in its early days. Creating reliable ML pipelines involves knowledge of distributed number crunching and database systems, Message Queues and Event Streaming, Cloud Architecture and Containers, Version Control and CI/CD Techniques. To support ad hoc querying of data, data models do have a bias towards what has happened in the past with respect to the data warehouse structures unlike the non-ML aspects where data models are designed based on the business knowledge of what could be queried, however these scenarios too require an understanding of concepts from distributed computing and architectural choices taken around the data storage and querying technology stack.

9.3. Machine Learning: An Overview

Machine learning is an interdisciplinary field of artificial intelligence that aims to create algorithms capable of automatically learning from data and making decisions when facing new and unseen situations. ML strives to discover underlying data patterns and rules, which can then be used, for instance, to build business models capable of predicting a company's future success. Recent technological advances have paved the road to the current ML development state. On one side, the rising availability of massive quantities of structured and unstructured data enables the training of more complex models. On the other side, the dramatic increase in computing resources has made possible the training of more complex model architectures that can learn richer data representations.

The scientific origin of ML is primarily founded in statistics but has exploited and created important contributions from other areas, such as computer science,

optimization, game theory, and cognitive science. Traditionally, ML has been understood as the study of statistical learning algorithms that learn from labeled experiences from the world, consequently requiring supervised signals for learning. In recent years, there has been an explosion in the development of new toolkits and methodologies that have made convenient the construction of ML systems for solving many application domains, such as computer vision, speech recognition, text understanding, and recommender systems, enabling the proliferation of the use of ML in many other research areas and application domains. Additionally, other key aspects in the widespread development of ML have resulted from the adoption of well-defined standardized evaluations that have guided progress in many of these fields, as well as the availability of powerful open-source libraries. Recent years have witnessed exciting new developments in ML: the ability to train very deep neural networks, combining increasingly larger amounts of labeled data and compute resources.

Fig 9.1: Building a Big Data Pipeline

9.4. Data Pipelines: Definition and Importance

Data Pipelines are at the heart of any Data Engineering endeavor, but outside of FineTech, they are often designed and developed with little attention to performance, security or mitigation of friction loss. They underpin the collection of data and its transformation into forms that can feed Analytics Engine Output, or novel formats that can be braided together and fed into Machine Learning pipelines, or Data Lakes, but are often not treated with the care and investment that other systems that provide enterprise competitive advantages. In consequence, the data which the ML models are built has issues, and the ML Models have issues, or the data that feeds decision-making has issues, or the translation of an analytics into dashboards or application changes has issues, leading to much of the promise that Data Engineering or ML models can provide never being realized.

To be fair, because pipelines tend to be quite "low-level" pieces of code and design, and because data must flow around for data engines to be useful, it is almost impossible for an Engineer to be cognizant of all the places where friction is lost. For example, a pipeline could apply filters that significantly reduce the data flowing into the front end for rendering, and Intelligence or Analytics Code could then run in a fraction the time; or, a pipeline could pre-fetch the historical data needed for calculations that inform a KPI displayed on the dashboard, significantly reduced time to render, but how do you implement monitoring or intelligent logic on the generating of the Historical requests that speeds them up, but don't slow down the experience of deep intelligence and analysis? These are the types of decisions that lead to a data driven enterprise, and that shine a light on good Data Engineering principles that make pipelines so important.

9.5. Building Effective Data Pipelines

Data pipelines are of crucial importance for enabling data-driven solutions within a fintech organization. Data pipelines provide a continuous flow of data from the data source to its destination, e.g., an analytic, ML training, or serving environment, in order to provide relevant data at the right time for relevant use cases. Pipeline building in data engineering is similar to software engineering where various components are pieced together in order to perform a series of pre-defined procedures. The industry's common understanding is that the concept of data pipelines is more concerned with rules and not control. A "pipeline" is defined as a series of data-processing steps or a long, thin structure, such as a pipe, through which liquids are carried from one place to another. The process of building a data pipeline usually involves designing a set of data transformations that convert raw data, coming from one or more sources, into a refined representation of the data, stored in some database or data warehouse.

We typically decompose a pipeline into three components: data input/export interfaces, a data transformation language/model, and data storage. When implementing a production data pipeline, non-functional requirements also matter a lot when it comes to operational requirements such as performance, accuracy, and reliability. Generally considered non-functional requirements are ease of use, scalability/flexibility, productivity, maintenance, replenishment, deployment, cost of ownership, and support. Implementing a production data pipeline typically also involves addressing and balancing these different, sometimes conflicting, requirements. In this chapter, we take a look at the most popular technologies in the analytics ecosystem that provide building blocks for pipeline components.

In a nutshell, we dive into the details of answering the following questions: How to get data into your analytics system? How to model and store and query artifacts produced by running pipeline operations? How to perform operations that transform data and implement the business logic of your analytics application? All those questions are addressed by component technologies that we describe in the remainder of this chapter, with a focus on questions 1 and 2.

9.5.1. Data Collection Techniques

The physical (or digital) universe is populated with endless avenues of data. Every action spoken and unspoken creates and leaves a trail of data for the collecting. The quality and quantity of relevant data directly correlate to the analytical business value of the actions and decisions made based upon those insights. We will highlight several techniques and ways to consider when developing data collection strategies suitable for vertical market financial technology products.

Many user interactions with fintech products are data-rich. Purchase and transaction details related to actions taken by users create data tables fitting for entries in the data pipeline. The details included in these transaction entries feed use cases for modelling with rules and exceptions. Enhancing product transactions with enriched data helps fuel product features requiring insights for decisions made. Incorporating rich identity and attribution data in real-time directly impacts analytic decision-making and associated results.

In addition to transaction data, user attributes and behaviors suggest to fintech product engineers and designers plentiful paths to consider for technology enablement. User portfolio attributes such as asset balances, credit scores, risk appetites, net worth estimates, investment preferences, goal preferences, and financial literacy increase relevant digital insights. Over time, understanding user digital engagement paths help product teams continually enrich the behavior data being collected. How long does a

user spend researching various components of a money transfer process? Which electronic devices are users employing to evaluate the particular feature? Do users navigate away from the fintech vendor to search for competitive product offerings or evaluations? When they return to evaluate their vendor's fintech product, how much time do they spend before making a decision and taking action?

9.5.2. Data Transformation Processes

The raw data collected through different data collection techniques is of little use in the original format. Data contains unwanted information such as duplicate, irrelevant, and extraneous data, which should be removed. Moreover, it needs to be organized in such a manner that it can be easily queried to support subsequent data query and analysis. The data organization and cleansing process is called data transformation. Data transformation allows businesses to create datasets that store useful information as opposed to raw data containing both useful and unwanted information. Data transformation consists of three distinct processes. The first process is data cleansing. Data cleansing supports the removal of unwanted raw data. Duplicate records should be removed, as they introduce duplication in analysis results. Moreover, missing values in relevant data attributes should be addressed. Missing values may include unknown, null, N/A, infinity, and blank values. The different methods employed to handle missing values include discarding records that contain missing values, or inputting values attributed in the missing fields. Records outline the same subject object or event, but have different values for different attributes. The second step in data transformation is standardization. Data standardization allows for the conversion of data values to a common scale, without distorting differences in the ranges of values. This allows for disparate data collected and stored in different scales, ranges, and accounting currencies to be compared and analyzed easily. The third and final step, known as data aggregation, allows data from different sources and in different data structures to be organized and stored in a single unified conceptual model. After data transformation is complete, the resulting datasets can be analyzed upon or queried using a query language. Business intelligence tools can be used to bring the transformative data to light, and useful, actionable insights can be derived.

9.5.3. Data Storage Solutions

Data Warehousing solutions are optimized for analysis and offer the advantage of reducing complexities for various reports needed by different LOBs. But ETL operations running on top of it can be scheduled only after business hours due to complex data models and increased data volume. Other than DW, the data is also being served out of

transactional apps like CRMs, ERPs, or even data from marketing automation tools. Each of these data sources has its own schema and transactional data served out of it is specific to the business domain. But this data also has to be merged with the ideas at various levels.

Modern businesses are moving towards a hybrid data storage architecture pattern in which they house their data in a data lake and ingest a lot of the data generated in an online transactional processing database and outbound analytics performed running analytics queries in the data lake. In such a hybrid architecture, the data lake houses raw data from multiple sources and is run over for analytics and reporting using tools. ELT patterns are preferred in data engineering.

Earlier, the backend to data engineering was either no-sql solutions or short-lived transactional apps on traditional OLTP databases. Lately, cloud-native solutions implement your Data Warehousing and Data Lakes storage in the cloud. For fast-moving analytics or reporting need, solutions crunch massive amounts of data quickly and power the reports.

9.6. Integrating Machine Learning into Fintech Products

Machine learning techniques can be used in fintech products that rely on traditional data analytics methods in order to address some key problems at a larger and faster scale in a more reliable manner. Predictive analytics, which consists of forecasting some future outcomes based on past data, allows fintechs to assess a mutable risk from a transactional activity based on a whole customer's historical momentum, and create suitable corrective actions or activate alerts. For such application, ML tools predict outcomes such as customer default, churn, upsell, cross-sell, early engagement, among others, based on classification techniques. Supervised ML methods can enhance predictive analytics leveraging different classification techniques. Regression or survival analyses can predict specific values, upon problem formulation. While structured data regarding internal company practices is sufficient to engender early customer behavior towards engagement, propensity, risk, and default prediction use cases, competitors' past behavior data from unstructured financial reports may boost top accuracy rates.

Risk assessment is one of the oldest machine learning use cases in fintechs and includes credit scoring models, anti-money laundering compliance, and property estimations for P2P lending marketplace. Fintechs may enhance credit scoring highly relying on maverick behaviors captured from the large traditional behavioral data tracked updating their characteristics in real time. In addition, some companies may choose to adopt and sanitize commercial or shared data that provide deep insights about reclusive credit users. Traditional financial and legal market data may be ameliorated with geo-location

clustered data for police checkpoints around an address, getting vast insights into formerly unnoticed suspicious criminal situations. After credit scoring, property suitability prediction is one of the oldest use cases in fintech.

9.6.1. Predictive Analytics in Finance

Predictive models based on Machine Learning (ML) have become a de facto standard for many fintech products. From apps designed to predict which stock will go up based on social media chatter to applications that rely on reinforcement learning (RL) to search for the best strategy on how to play a video game, ML is ubiquitous in finance. But it was not always that way. Research in finance has had you could say a classic resume, where econometric models were the dominating paradigm and ML was considered too stupid or naive. The introduction of large datasets with the internet boom, as well as the availability of faster computers, allowed for feature-rich ML algorithms to become competitive with their classic counterparts in the 2000 decade. Between the decades of 2000 and 2010, many Wall Street firms started employing ML-based products. Since they were either the only user of the strategy or had a first mover advantage, these strategies were either very profitable or difficult to replicate.

Since 2010, training a ML model has become one of those tasks that do not require an army of PhDs. The growth of the "as a service" model has given birth to an army of copycat ML products that try to predict every possible variable out there by scraping large amounts of public data. However, the battle has just started. Traditional econometricians have come back and are trying to make the best of ML in finance by trying to build prediction tools that are simpler, take less resources, and avoid overfitting as much as possible. These users dive into concepts like causal inference, generative models and prediction intervals. In the next sections, we will introduce the application of ML into credit scoring systems and transaction monitoring systems, two of the crown jewels in non-capital market based finance. In them, we will aim to compare the ML-toe versus the traditional econometrician to show which tools are used by the respective camps.

9.6.2. Risk Assessment Models

While credit scoring was perhaps the first ML application in banking, it seems that loan approval decisions are becoming more frequently automated by ML engines today. Banks rationally decide to invest in ML models to predict the probability of default because with accurate credit scores, they can effectively serve their retail client base and offer large loan books at reasonably high interest rates. Loans, both for individuals and businesses, are a major source of profit for banks and related institutions, so being able

to optimize risks and returns are a huge push factor for ML adoption. Additionally, more unsecured lending is being given, including mortgages and home equity lines. Also, there are near banks that provide unsecured loans for small and medium enterprises using simplified KYC requirements, employing social media analytics, and using ML and AI decision systems including verification methods that are quicker and use open banking. For these companies, the simplicity and scarcity of transaction data could either drop their performance or make it too hard to compare with the scores of other institutions. This prompts firms to use adaptive heuristics and/or local comparisons. Most of the above companies will only be able to optimize the adverse selection risk element of credit risk, while there are parameters that come from different sources and economic aspects of portfolio default probabilities, primary, as well as obligor-specific correlations, availability of safety margins, default variations and other systemic risk conditions that ultimately shape credit spreads. It is well known that traditional credit scores sometimes fall behind in providing the correct signals when other economic variables change.

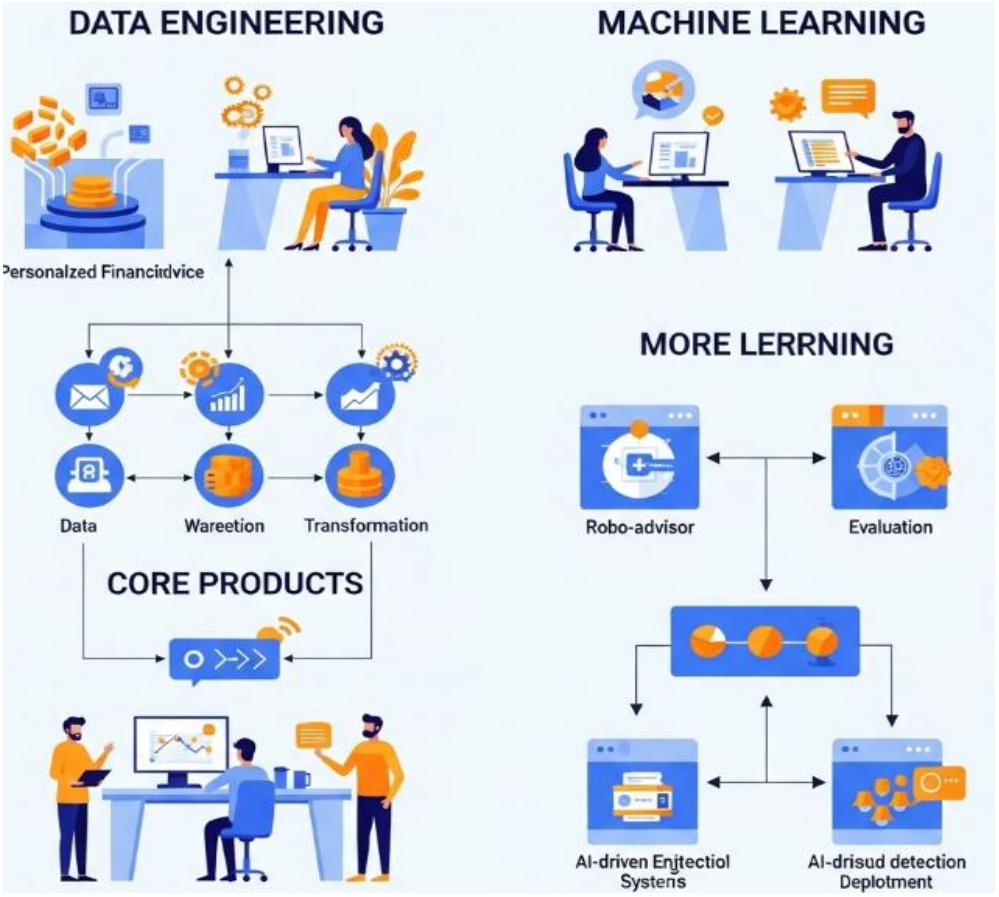

Fig 9.2: Fintech with Data Engineering

9.6.3. Fraud Detection Systems

We must always have to combine current machine learning techniques with the current storylines of potential fraudsters to enhance the efficiency of banking fraud detection systems. This feedback loop is the key to allowing these automated systems to predict and prevent the big data, big problems of financial fraud. Internet and mobile banking systems offer banks a chance to expand their client base, to improve relationships with customers, and to boost their revenues, but there's one downside: there's also big data, big problems to contend with. Online banking systems are seen as a safe haven, a treasure chest for organized cyber-crime on a global scale. Only the banks with automated fraud monitoring systems in place are able to flag potentially suspicious transactions for manual review, given the sheer volume of transactions at high-traffic banks. How to have the best possible fraud detection tool? An ideal initially would include fully-automated systems continuously scanning for suspected fraud, user-friendly software tools on the bank's side allowing detection staff to flag and monitor suspicious customers, and an early warning fraud alert system. Moreover, non-personal data also plays a strategic part here; together with rules set by security experts it helps tune the system in order to better minimize fraudulent transactions without bothering customers with unnecessary alerts.

All banks today have in place some automated procedures aimed at tracking the progress of transactions in real time for any account that might seem to be involved in fraudulent activity. Automated systems can stop a huge number of potentially fraudulent transactions; however they can also alarm on transactions that afterward may be found to be authentic, making a slight risk of "false alarms" at the customer's expense. Maintaining the accuracy of fraud detection systems is key to preventing lots of dirty business and to preventing lots of upset customers. To accomplish this last task, financial institutions continually adjust the algorithms which power their fraud detection solutions every time the procedures identify a suspicious case.

9.7. Real-Time Data Processing

Typically, we consider batch modes of operation, in which we process groups and sequences of transactions and logs at specific time intervals. The resulting delays usually range from seconds to minutes. In many applications, this is not acceptable. Think for example of an attack against a fraud detection service that floods it with random events with a known invalid status. Such an attack is usually detected by the service itself within a few minutes, but the temporal proximity between the different events makes it possible for a malicious user to execute it at a very fast pace without any chance of being detected for this transaction — just because he knows launches lots transactions with the same attributes but at least two seconds apart. Real-time data processing for fraud detection

transforms this seconds-to-minutes window into milliseconds, alerting a company as soon as a user triggered a previously unseen transaction.

This push towards lower and lower latency times in the detection of certain types of events correlates well with the increasing speed at which we generate data. From a couple of years ago, traditional databases started being supplanted in certain applications by NoSQL solutions that departed from the traditional model of ACID transactions that limited throughput significantly. Incorporating models such as columnar storage, eventual consistency with domain event sourcing, high availability and partitioning, and document-oriented models, these solutions support huge traffic, with writes, deletes and updates in the order of terabytes per hour.

In this context, stream processing frameworks were developed to offer real-time data processing capabilities, allowing users with a specific class of applications and use cases to harness the inherent speed of NoSQL database engines for fast alarms while enhancing the use of resources involved. Stream processing typically uses efficient FIFO buffered queues in the typical publish/subscriber data flow pattern.

9.7.1. Stream Processing Frameworks

Over the last decade, a number of big data engines have incorporated streaming. For example, one engine has added micro-batch processing capability and is in the process of converting its engine into a generic engine that accommodates both batch and streaming processing. Another is a distributed stream processing engine that scales well for both batch and streaming workloads, and offers good support for low-latency processing. A real-time stream processing system is also available. Based on certain technologies, another engine uses the same mechanisms for fault tolerance and isolation. However, one can be thought of as a stream classification engine rather than a full-fledged stream processing framework. Another engine takes an application centric view of stream processing. A programming model, implementation, and service for stream processing is also provided, based on the concepts of pipelines, which convert input data into output data, with the help of transformations that, in turn, use user-defined functions to process data. One system can take both batch functions and streaming functions as inputs for constructing batch and streaming pipelines. Its programming model is similar to another model and internally is built on top of a specific framework.

Another service has a SQL-like language for stream processing. These frameworks also provide a large ecosystem for building job management and operations infrastructure for stream processing jobs. One organization, focused on using a specific streaming engine for batch and stream processing, has developed control plane and data plane tools and experts that are focused on building an agile and performant infrastructure in which

business stakeholders can write queries as prose and a specific query language, which will be elastic query optimized and batched or streamed according to their specifications with minimal delay. This system provides visibility into how these complex dependency-laden jobs are operating and the ability to get alerts when things go wrong.

9.7.2. Batch vs. Stream Processing

Batch processing and stream processing are the two types of processing for computation. In batch processing, data is collected over a fixed time period, and when a sufficiently large batch is accumulated, computation is done over it. Batch processing has been the mainstay for processing data at scale for the last 30 years. Its directly available programming model over both data and its transformation has made programming straightforward. Tools have made it possible to process large-scale data across large clusters at low cost.

However, the total time taken to wait for data to arrive, finish a large computation, and learn from it, is too much to enable interactive-type learning. Some applications and use cases require a low latency computation where each data point as and when it arrives needs to be acted upon, quickly enough that the user sees a meaningful result. In these cases, batch processing is not helpful. Startups helped pioneer new types of large-scale data processing to provide real-time information to users, using stream processing systems. Their work continues to drive how stream processing is done today, helping search engines and news aggregators provide fresh information to users.

These systems were preceded by a myriad of stream processing systems built using different types of computation and storage systems. Pioneered the Map-Reduce model, following it with the Map-Reduce-Online model and File System to handle at-scale batch mode processing. Developed their internal architectures along these lines. Related work in academia produced systems. Realized and expanded on the real-time requirements of their businesses, through the system. Soon followed through with a clearer public revealed architecture, the system.

9.8. Case Studies of Successful Fintech Implementations

Across multiple facets of finance, business, and technology, only a select few have successfully merged their strengths to build innovative products. Specifically, these few implementations embody the use of advanced data engineering pipelines and machine learning investments to create and drive growth for such products. In this section, we cover a few such implementations from the Fintech sector that have successfully leveraged machine learning and data engineering at their core.

The first example comes from the world of data-driven advisory services. A large financial institution built an in-house engine that sources alternative data from various points to gain a better insight into customers' profiles and activities. Using this data, the team built an internal model to score clients using an insight-score metric based on their algorithmic input. By developing a personalized interaction with clients based on the scores, this institution successfully enhanced customer engagement and ultimately improved customer retention in a high-churn environment. The solution was able to predict likelihood of purchase and allowed the bank to tailor its marketing accordingly. This infrastructure for predictive insight-scoring provided a major competitive advantage to this institution when compared to other financial services firms who relied solely on traditional methods for such scoring.

The second example comes from the world of robo-advisors. Traditionally, investment management has always been a service reserved for high-net-worth individuals. Even that investment management space has remained archaic and traditional, providing limited technology-enabled tools to assist in making what is at the end of the day a technology-related business decision. Specifically, investments are long stayed with the traditional means of portfolio company valuation utilizing fundamental analysis based on value biases. It has been challenging for non-high-net-worth individuals to build wealth using passive investment strategies due to cost of portfolio management. However, the recent emergence of online investment management services that utilize technology has disrupted this long-stay investment management business.

9.8.1. Example 1: Credit Scoring Models

An area where fintech companies have been particularly successful in implementing data products is credit scoring models. These have been around for decades, but fintech has transformed credit decisioning by bringing greater speed, advanced data sources, and improved modelling tools. The earliest modern scoring models date back to the 1950s when cash flow information was used to segment credit applicants into risk categories for banks, enabling internal models to be replaced with automated tools. This model became increasingly sophisticated over time, and it soon became the industry default for retail consumer credit scoring. Using past defaulting behavior of large pools of borrowers, statistics were pioneered to build models that predict the likelihood that a borrower would default on a given loan during a specified period. The model was similar in principle to existing scores, but a big benefit of the statistical approach was generating a customizable model that different lenders could tweak with rich data sets from their own past transactions to optimize predictive accuracy.

Since then, many more data sources (both traditional and alternative) and modeling tools have emerged. But traditional lenders and startup fintechs alike have still relied primarily

on the data-agnostic tools pioneered in the past. Existing fintech lenders have been disrupted by upstart companies that have launched next-generation scoring models that combine painstakingly designed data features with the most advanced AI modelling techniques. This has produced scoring models that consistently beat competing models built using legacy data sources, feature sets, and modelling tools. These new models have been able to provide consumers enhanced access to credit at lower interest rates.

9.8.2. Example 2: Personalized Financial Services

Digital-native small to mid-sized companies (SMEs) are leading towards the world of personalized and innovative financial services. They are seeking neobanks for transparent operations, banks with ready onboarding, seamless integrations, competitive pricing, and most importantly empathy. A pioneer in streaming analytics, leading story-driven personalization for small business financial services, launched a personalized content experience. Actively engaged users are 2-3 times more likely to convert on an offer, add tools to their accounts and come back to complete the online process – perhaps the highest engagement ever seen in any industry for a digital-first small business target. The user interfaces are like sensors picking up user interest and engagement based on the user journey, emerging customer data streams. The user journey is mapped inductively and used to create interactive campaigns around highly personalized calendar events. Dynamic monthly financial reports automatically prepare users for their accounting. A cash flow report showing pizza or personal expenses drives acceleration of tool adoption by pizza restaurants and real estate companies.

The company's advanced predictive analytics engine constantly sifts through customer data streams to identify the likelihood of their acquiring a product for each seamless data push. The products being presented at the appropriate time through the correct channel have been proven to have the highest acceptance probability. This enables banks to allocate a significant portion of their marketing budget to these funnels, increasing onboarding speed multi-fold and increasing the chance of proper ongoing advising: Customer lifetime value, and interchange fees staying true to the core business. By using machine learning and prediction, banks and fintechs will be able to concentrate naming a pain point, be it invoice or expense automation, and ensure that the solution is delivered at the right time and state.

9.8.3. Example 3: Investment Management Solutions

The current investment management solutions rely too much on past price trends and price predictions to assess how good a security is, as if that alone is enough to invest or select or rank securities by attractiveness or to estimate a one-to-one asset price. There

is also too much focus on high-risk high-returns and absence of concern for highly reputable organizations that provide consistent lower, yet no-present-risk returns to their continual investors. Furthermore, the current ratings are static in nature and don't consider many market factors for returns on assets in multiple market contexts for one-over-the-other relative ranking. They rely too much on assumptions instead of scientifically validating the assumptions beforehand and learning how much those matter for outcomes. These assumptions are themselves takeaways from preliminary exploration, to which data process pipelines and data models can be applied.

The investment management pipelines have recently blossomed with the asset level data provided by the market as well as the availability of learning algorithm pipelines. The main ingredients for the learning pipeline are making better, more science and experience-based data features, tuning appropriate models after mapping the output into either a ranking task or a reversion to mean prediction for asset return levels, and considering asset prices from an allocation or portfolio context for multi-arms bandit problems or any other means-based statistics for a batch of assets. The output of the investment management pipeline may take different forms, depending on which of the task objectives were considered while training and tuning models. The output could be dynamic ratings or a recommendation of active allocation weights either optimally or by budgeting relative to expected performances in different contexts, or it could also be a set of prohibitive lower or upper bounding annuity ranges with respect to risks for a defined investment duration and timeline.

9.9. Challenges in Data Engineering for Fintech

Real world data, unlike ideal data models, present considerable challenges to fintech data engineers. First, real world data size can be enormous - which means that, particularly in the world of big data pipelines, engineers have to solve the problems of scalability - processing and storing big data. Second, real world data, particularly big data, is messy - socio-economic data is often crowdsourced and may contain erroneous data points that can affect downstream analytics: not just anomalies, but ethical concerns about bias and faulty classifiers. Data engineering pipelines may additionally have to support model, concept drift and real-time processing. Third, the world of financial technology is heavily regulated, with strict industry and government compliance rules for customer and enterprise data. Data engineering teams must guarantee that models are interpretable; that data underpinning models is cleaned and labeled; that customer privacy is ensured; that any compliance issues are resolved.

In Fintech, issues of data privacy and security with sensitive individual and enterprise data are paramount. Technical measures such as encryption, honeypots, detection and response systems, as well as data minimization, have to be implemented, and it is usually

a considerable challenge for engineers to ensure that only the minimum amount of required information is retained. Having good identity and access management software tools, together with discrete physical and other office controls to limit access to sensitive data is equally important. Protecting intellectual property, preventing exfiltration and attack, and maintaining the integrity of systems are equally challenges, as are actually using all these preventive measures.

Good fintech products typically need to have the capability to efficiently process and store large amounts of real world data, and have scalable data management pipelines. Specialized tools are needed to address such scalability issues. Therefore; problems such as streaming, sharding, batch processing must be addressed before the data is made ingestible for datastores. Also; storage APIs need to be efficient not just for upload (batch) but also download. Crosscloud ingest needs to be fast and require a parallel build pipeline processing. Building a managed services abstraction around the ingestion and transformation process is valuable for a smaller company.

9.9.1. Data Privacy and Security

Financial institutions have had to secure sensitive data well before the internet. And now with opportunities to leverage user data with novel and performant products, it is only fair that users should feel empowered to control their own data. The current challenge associated with that is to provide data security in a performant way. It is very easy to build barriers to access information by burdening user experience; the goal is to allow user access as transparently and securely as possible. While many fintech companies spend time and technology resources building advanced capabilities to perform authentication, data governance and monitoring, user experience at the data edge is still cumbersome, simply to comply with regulators.

User authentication at the data level can be cumbersome. There are novel ways that have become available to fintech companies to protect data while not compromising on performance and while creating an emulated transparent access. Compound that with the fact that edge location systems have made distribution of data faster than even before – i.e. the user does not need to travel long distances to get their data – and the use case is ripe for experimentation with cost at the requirement of developers to not compromise on user experience. Fintechs can explore services available in the market today that allow for organization capability to encrypt data at rest and at transit at their own chosen unique key that are only accessible for the personnel who need to see that data at the firm. These services can provide APIs akin to traditional volume storage and allow for point-in-time access to ensure the level of transparency required without compromising on encryption capabilities – and at affordable costs too.

9.9.2. Scalability Issues

The rapid growth in the volume and variety of financial data has outpaced the capacity of existing infrastructure capabilities for many organizations. In fintech, particularly for larger platforms, fast data streaming, low-latency access, and real-time decision-making capabilities are critical. This has created a renewed focus on the challenge of scaling solutions to meet consumer needs. Strategies that have been employed include more focused data-collecting initiatives, enhanced exploration, increased cleaning and relying on enrichment, built circular data pipelines, unified enterprise operations, optimized data storage, employed a data model that supports varying structures, integrated external data, behavior modeling, and investments in transactional systems and infrastructure.

For many providers, the issue is compounded by the fact that many current systems may have originally been designed for archival storage purposes rather than for the needs of fast data streaming, real-time data access, and low-latency data processing. Many financial organizations are accelerating their technical debt confrontation. By switching to new cloud-based solutions that can allow for increased scaling and more flexibility and adaptability, they can be better positioned to help their organizations optimize business value moving forward. Scalable data engineering operations enable collaborations across cross-functional teams, decrease risk of project failure, reduce time to market, implement enterprise best practices, enable evolving and emerging technologies, apply robust data governance, and enhance data asset value and contribution to the overall organization.

9.9.3. Regulatory Compliance

Fintech companies and other service providers must strictly follow financial regulations. It is easier to create applications than to validate whether there is any regulation that affects it, because even a small function of the application interacts with regulatory functions or has regulatory data. Applications have to be validated with different kinds of regulations, which are constantly changing, at different times by the companies and by different governing bodies. This means that there will be changes that might affect the use case. There might be exposure to tactics to circumvent the regulations via loopholes and the responsibility lies at the door of the companies or parties connected to the finance sector.

New use cases requiring the use of solutions will require extensive consultation for the scope of the application, and still there might be disputes on the depth of coverage during its application as regulations are volatile in nature with differing enforcement guidelines across geographical regions. Complying with regulations costs a significant amount of money for those obliged; Noncompliance has its consequence and can lead to hefty

regulatory fines. Continuous monitoring, reporting, validation, and auditing are essential for all actors in order to help in the smooth functioning of the system. Thus, all violations should be detected early enough in order to take countermeasures, mitigate personal and institutional risk, and avoid incurring losses. Existing strategies primarily leverage decision trees or rules, goal-based membership function, knowledge graphs, reinforcement learning, and models to enhance accuracy, explainability, and detect harmful events in real time.

9.10. Future Trends in Fintech and Data Engineering

The field of Fintech and Data Engineering is rapidly evolving, with new developments emerging regularly. In this section, you will learn about three upcoming trends in Fintech and Data Engineering. This includes AI automation of finance, Blockchain, and open banking.

As companies have experienced firsthand, the attractiveness of personalization is too strong to overlook. Our expectations have been set high by these data-drenched behemoths. Fueling their trajectories in personalization and recommendation is AI. Today it is reinforced with machine learning-driven recommendations that tailor experiences to individual users. Consequently, it is widely anticipated that AI will fundamentally affect how we interact and perceive all things financial. Entire processes and workflows are expected to be automated, freeing employees from repetitive manual work. Data engineering will, inevitably, be at the heart of these initiatives. As data pipelines augment the curation of data through which machine learning experiments and models are built, clean, usable datasets will flow into model devops tools that automate the model building and operationalization process. This, in turn, will drive the democratization of which financial functions are automated.

When we think of these resource-intensive self-driving cars, increasingly resembling a mobile version of a data center, the idea of a bank in every device is likely to sound like science fiction. But advanced fintech applications are poised to lead us to a world where everyday appliances will be functional, secure, and transparent payment interfaces. We are likely to see Blockchains paving the way for creative new types of business models that foster deep integration between financial and data engineering functions that go beyond information transfer. Smart contracts encoded on Blockchains are expected to support the conviction that the role of banks will shift from being trust brokers to become trust sponsors at the heart of business ecosystems. Curb-friendly standards, superapps, and the open ecosystem approach are likely to be the common blueprint in a post-granular world for data-driven products and services that will ultimately reshape banking as we know it today.

9.10.1. AI and Automation in Finance

Fintech companies are at the cutting edge of leveraging new technologies to transform the way people manage their finances. The combination of technology and finance is nothing new. The history of finance is connected by constant disruption and innovation to help us manage our businesses and lives better, and fintech is the latest innovation in this regard. What is new is the speed at which these changes take place. With near zero transaction costs, the rise of global platforms, the fintech revolution is solving hard unsolved problems at a staggering pace.

Artificial Intelligence (AI) advice on managing finances is just the latest in a long line of financial advisors, but worries about scams and lack of humanity in financial decisions have delayed our trust in using machines to help manage our money. Other areas in Finance have had a much easier time in recent years, for example, predictive models in treasury offices or sales forecasting models predicting revenues are used on a daily basis to guide decisions by large financial institutions. Machines are now a critical source of advice and give banks a significant competitive edge. Large financial institutions use predictive models to help improve marketing, increase sales, and manage attrition for retail banking. Automatic customer scoring is used by commercial banks to prevent loan defaults. Banks have created digital assistants, chatbots, and voicebots to improve the customer experience. Are all of these applications AI?

9.10.2. Blockchain and Data Integrity

As fintech continues to disrupt and innovate the finance industry, traditional processes are being challenged and entire new areas are being explored. One geographical area poised for tremendous growth and innovation in fintech is the South East Asia region which has a combination of endemic challenges and an explosive growth stage where innovative fintech solutions can be rapidly deployed and adopted at scale. Another area that is booming at this stage is Non-Fungible Tokens and the evolution of the metaverse. Companies committed to build the metaverse, whether that relates to gaming, retail or social engagement need specialized fintech solutions as large volumes of transactions and assets are being built.

The blockchain has generated a disproportional amount of interest and hype as companies in a variety of industries start exploring the use of blockchain technology as a strategic capability. The defining characteristic of blockchain technology is the immutability and resistance to fraud since for a transaction to be fraudulent, a hacker would have to hack a number of participants and that would quickly collapse the integrity of the blockchain. At its core, the blockchain is a distributed ledger where every participant has a copy of the ledger which records a history of transactions. Ledger

systems allowing read-write capabilities can become susceptible to a variety of integrity attacks including unauthorized modifications of information, insertion of false information not supported by any transactions, and deletion of transactions.

9.10.3. The Rise of Open Banking

Recent years have observed the increasing emergence of regulations around open APIs from financial institutions. Essentially, these APIs allow third parties to connect, obtain bank account details, facilitate payments, and offer services that were necessarily confined to the region of the financial institution itself. Still, there are security boundaries introduced to ensure that only trusted service partners can access user account details and execute remote transactions on the user's behalf. The most emblematic regulation in the area of Open Banking was the one introduced by the European Directive.

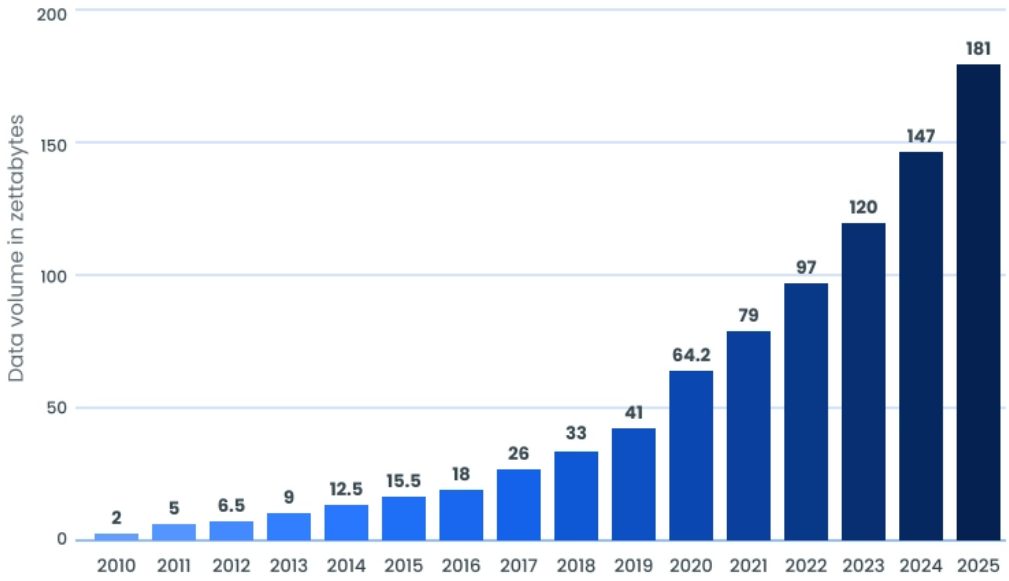

Fig : Data Engineering Trends

The basic concept of open banking is to allow a bank (or any other payment service provider linked to a bank) to share their bank account data with an authorized third-party provider through the implementation of a Secure Customer Authentication process. Bank account holders could then allow other service providers to access their bank account details, like balance amount or transaction details, and to execute payment transactions on their behalf. Open banking APIs thus increase customer business opportunities for fintech companies, allowing them to offer new products or services that would not typically be possible only around their financial institution without a bank partnership.

It is not necessary to highlight that the promulgation of security and trust features around banking APIs was a necessary step for Open Banking to exist. Customer protection through transaction verification and message security on the side of all consumers, fintech services, and banks is likely to be an ongoing area of innovation in the payments and banking industry in the near future. Thus, future fintech product/service innovation is likely to come from trusted API-based exchanges of financial account data between banks and authorized third parties in the area of Open Banking, pushing the necessity for banks to continuously innovate their core business, as customer loyalty cannot be taken for granted anymore.

9.11. Best Practices for Developing Fintech Products

Building a new financial technology (fintech) product from scratch is not a trivial task. Financial markets are heavily regulated, and customer sensitive information is usually involved with many fintech startups. This often creates significant entry barriers in terms of costs and risks to product development. In this chapter, we cover some of the best practices involved in the product and business aspects of building fintech products.

The fintech sector has seen the explosion of technology-enabled financial products that make users' lives faster and easier. The overwhelming majority of these products leverage modern cloud-based infrastructure to remove the burden of setup, security and technology maintenance from the end user. There has also been a rapid proliferation of companies focused on the development and management of fintech products or providing tailored solutions for specific financial products. Information and ideas move faster than ever, and almost anyone can enter the game, so the rules to success are changing.

To be successful, fintech products must be easy to set up and easy to manage. Following general industry product development best practices and methodology that facilitate a high-quality, quick turnaround of product can significantly increase the likelihood of being the leading innovation in a fast-paced environment, and also help to scale product to support increasing demand. For financial products, agile development methodologies and facilitating a user-centered design experience are two particularly important areas when considering product development best practices. A strong alignment with customers and their business goals is key to building a successful product.

9.11.1. Agile Development Methodologies

The rapid change in our lives demands the creation of new Fintech Products in brief periods, along with official approval processes which present both challenges and

opportunities in a regular end-to-end agile lifecycle, because each product involves changes in a regulatory framework which should be aligned with a special product. The agile methodology was created in the largest software houses for software development or software-based products, but not for Fintech or finance products in general with a very close dependency with the regulatory area. Nevertheless, the agile philosophy should be adopted with acceptance, sometimes overcoming regulatory authorities' concerns, in various stages of a Fintech product lifecycle. Agile principles can help Fintech companies to identify the specific Fintech product stakes and agree on an adapted agile methodology capable of creating the product and increasing the impact in case of hiccups.

Fintech companies use the standard process during the whole Fintech product Idea phase, in which multiple Fintech ideas are analyzed, with exploration scorecards and problem-solution validations. The result of this phase is the creation of a certain number of project proposals with a justified exploratory phase conclusion, which are submitted to the financial authority for validation and are then ranked according to a maximum scorecard model. The selected projects will then enter an exploratory phase, focusing on short validation cycles during which each Fintech company will define the first prototype, the exploratory phase objectives, the product prototype architecture, the work packages/stories to be validated, and the exploratory phase deliverables including proof of concepts.

9.11.2. User-Centric Design Principles

Fintech products have a unique set of characteristics that differentiate them from products in any other industry sector. In addition to being regulated by authorities, fintech products are the target of the scrutiny of multiple stakeholders: banks and financial institutions who partner with the fintech to provide infrastructure; users who have no previous experience with its kind of product; defenders of transparency who, in practice, are aligned with authorities and regulators; authorities and regulators. Product characteristics like pricing and security are more scrutinized and tend to be more important at making a decision than in other non-fintech products.

Also, fintech products are not the result of a prototype-development cycle, but rather of existence-critical activities that if not developed in the first place can cause mistrust and skepticism. This is because fintech products are relatively new and restrictive, being disseminated among users. Thus, the users of early stage versions are that initial group of people who will determine how and if the product will be successful. The demands and expectation of this group are completely different. For this reason, user-centric methods, like design thinking, should be employed as early as the pre-development stage. Design thinking emphasizes serendipity and radical ideas, being less directive and

prescriptive than current Agile methods. It encourages trial-and-error experimentation more than implementation short deadlines and constant delivery of products, and the involvement of the users or some part of them is done much earlier than usual. Then the initial demands and expectations of the users can be matched. At least these methods can influence and guide decisions that will either succeed or fail in attracting the users of a Fintech product.

9.12. Conclusion

In this chapter, we compared traditional Financial Technology (Fintech) processes and proposals with novel Data Engineering and Machine Learning (ML) Pipeline use cases. The Fintech processes and proposals were discussed with a thorough approach, exploring key areas of Fintech's emergence, its innovative proposals, and its challenges. The data pipeline examples were explored in longitudinal Cohort Analysis and Fraud Modeling, as in other supervised and unsupervised examples.

In summary, these Data Engineering pipelines increase the availability and quality of Fintech products and provide a reasonable solution to Fintech shortcomings in transparency, inclusion, and automation. With practical use cases in mature ecosystem countries, we validated the concepts by peer Financial Market experts and additional research scope. Several market players are growing with different challenges for short term offers with increased capabilities and transparency. With the long-term evolution, it expects to have a multi-market ecosystem with cooperation and competition roaming Fintech challenges with better products to final clients and society well-being.

We believe that, in the future, real-time Machine Learning capabilities will allow Financial Market main players to add better risk management, compliance, and client engagement in their proposals - relying on Fintech specialized players' disruptive creativity - creating a more robust system for companies and consumers alike. We agree with other researchers that state that the current rapid pace of non-linear Fintech ecosystem changes is increasingly reaching and evolving traditional players' compliance and risk management challenges; therefore, Fintech partners' establishment is essential as market offers evolve.siness model to drive change and, in some cases, perform a fundamental reshaping of financial services.

Yet despite increased levels of global competition and a number of consumer-friendly product changes, broadly speaking, the efficiency and profitability of the traditional banking sector worldwide has remained largely stagnant. Recent surveys among central bank executives considered the increased use of fintech services in the provision of banking activities in most G20 nations to be an important opportunity.

References:

Chen, M., Mao, S., & Liu, Y. (2014). Big Data: A survey. Mobile Networks and Applications, 19(2), 171–209. https://doi.org/10.1007/s11036-013-0489-0

Kelleher, J. D., Mac Carthy, M., & Korvir, D. (2015). Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies. MIT Press.

Gai, K., Qiu, M., & Sun, X. (2018). A survey on FinTech. Journal of Network and Computer Applications, 103, 262–273. https://doi.org/10.1016/j.jnca.2017.10.011

Akhgar, B., & Sutan, A. (2021). Artificial Intelligence and Big Data Analytics for Smart Financial Services. Springer

Breck, E., Cai, S., Nielsen, E., Salib, M., & Sculley, D. (2017). The ML test score: A rubric for ML production readiness and technical debt reduction. In Proceedings of the 2017 IEEE International Conference on Big Data (pp. 1123–1132). IEEE. https://doi.org/10.1109/BigData.2017.8258062