Deep Science Publishing
https://doi.org/10.70593/978-93-49910-41-6_11

DeepScience
Open Access Books

Chapter 11: Scalable artificial intelligence infrastructure: building tech stacks for financial institutions

11.1. Introduction to AI in Finance

Over the past few years, there has been a growing interest in AI in the financial industry. AI has shown great promise in enhancing existing systems and creating new insights in a wide range of applications, such as algorithm-based trading, fraud detection, risk assessment, and client services. Despite this, most financial institutions are still in the early stage of AI adoption. While they are executing pilot projects, they are facing the challenge of deploying these models due to infrastructure and compliance issues across the full value chain. We propose a full-stack distributed machine learning platform that can enhance the efficiency of researchers and data engineers and provide tools to handle the heavy task of DevOps (Aversa et al., 2018; Ghosh & Albrecht, 2020; Hameed & Yang, 2020).

Our AI infrastructure is designed to be multi-purpose: (1) data and model governance, to ensure that all in-house models remain in compliance; (2) performance and cost efficiency, particularly for deep learning inference; (3) scalable orchestration on data-parallel and model-parallel with distributed deep learning for research and DevOps tasks. In recent years, the research community has proposed many system designs as well as software tools and libraries to help with these tasks. We recognize that there are several key design considerations for crafting practical software, especially for the heavily regulated financial services. We discuss our design principles and the design choices. To validate our design, we have successfully deployed our full-stack distributed machine learning platform in one of the largest financial organizations(Kshetri, 2017; Zhang & Li, 2019).

11.1.1. Overview of Artificial Intelligence Applications in the Financial Sector

Artificial intelligence (AI) and machine learning (ML) are expected to significantly impact the financial industry. The applications of these technologies include a wide range of scenarios such as chatbots, fraud detection, risk analysis, algorithmic trading, and financial recommendation systems. Many financial companies have been actively deploying AI and ML systems, but there is still an insufficient amount of research on methodologies for large-scale systems that can process massive amounts of real customer data.

One of the reasons for the insufficient establishment of a research methodology is the lack of large-scale, real-world AI system examples. Although AI and ML research uses numerous public datasets, most of these are too sanitized to be used for large-scale commercial products. As a result, financial companies need AI infrastructures with real-world capabilities that can process large-scale real customer data and serve a wide range of business needs. This overview provides an overview of challenges in commercializing AI and ML in financial institutions, introduces an end-to-end AI infrastructure platform developed for industrial use to overcome these challenges, and presents case studies of profitable AI services for financial institutions deployed using this platform

.

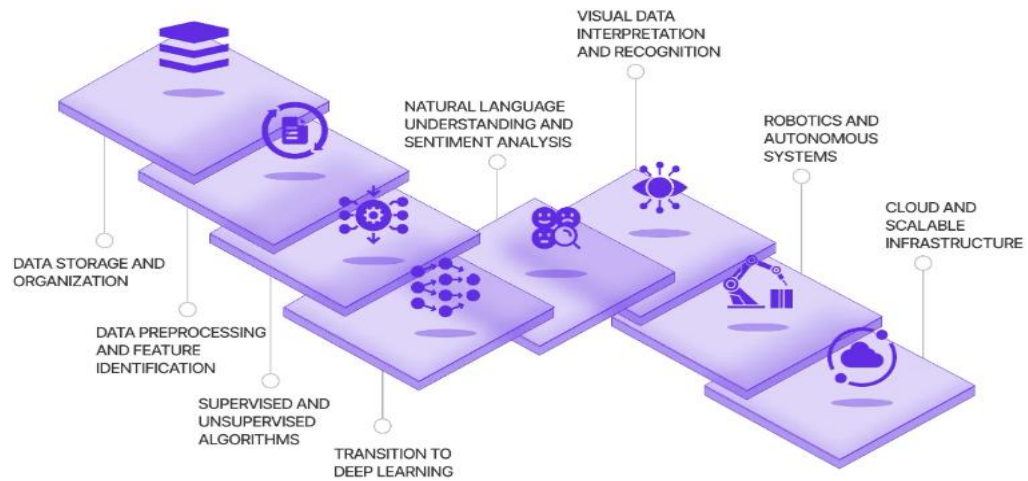

Fig 11 . 1 : AI Tech Stack

11.2. The Importance of Scalable Infrastructure

AI is unleashing a new wave of innovation. However, AI-infused applications can swamp network bandwidth, consume most of an organization's computing cycles, create extra load for storage systems, and burden IT staff. The core thesis of this paper is that

these negative impacts can be avoided by building the infrastructure to host AI effectively. Successful firms are transforming to be AI-infrastructure savvy and will then leverage the power of AI to drive their businesses. Today's finance industry is facing data management issues directly from AI applications. Whether they address a data-rich problem in their transaction business, try to improve the integrity of a high-frequency trading environment, reduce risk through imprudent management, supervise financial activity, or enhance the customer experience through personalized products, actors in the financial services have turned to AI solutions. This is now becoming the norm, and the next wave of competition will focus on creating an infrastructure aware of AI. Those companies that can train and host AI will win by taking full advantage of fact-based market decision-making. This paper is about helping financial institutions understand the importance of creating infrastructure that scales AI. It outlines the shape and the benefits that can be expected from infrastructure that can handle the growth of AI's phenomenal data sets, machine learning models, and associated management problems. It offers recommendations on how to implement the infrastructure. These pointers are designed to help integrate AI across critical business functions from trading to investment analytics, trade surveillance, risk management, credit lending, and customized financial services. This paper also touches on threats to success and shows how using articulated strategies can de-risk these illusory elements. With the AI market, a competitive landscape estimates to reach $55 billion in 2025. There is no surprise in this. AI is focused on manipulating sophisticated numerical challenges, of which there is no deficit in the financial industry. The motivation behind intelligent enterprises identifies AI as a critical operation. It should, after all, facilitate better decision-making, the identification of options, and an impressive ability to work with large volumes of data. However, AI is not just about choosing the right algorithm; it has to do with the practical and scalable application of AI. Instead, understanding AI infrastructure is important: the strategy that is usually challenging to quantify, sometimes in secondary areas. The aim of this paper is to help clarify the meaning of the concept and to help financial institutions understand that their next competitive frontier is to embrace infrastructure for scalable AI. It describes the relevant infrastructure, its advantages, the importance of the AI mission, and the challenges to the popularization of scalable AI infrastructure.

11.2.1. Building a Robust and Adaptive AI Infrastructure

The AI infrastructure is the overall system that supports the lifecycle of AI services. The lifecycle includes creating, deploying, running, updating, and monitoring AI services. Scalability usually means the ability to handle bursty traffic, sudden increases or decreases in throughput. For legacy systems that have been serving for tens of years, the requirements are evolving. Since it is very expensive and badly needed, we need a robust

and adaptive AI infrastructure that delivers the same level of services as expected and also copes with the continued evolving business requirements. The AI infrastructure includes the hardware infrastructure, such as the clusters, nodes, operations, or storage, and the software stack, such as virtualization, orchestration, and management platforms.

Model deployment and all the steps leading up to the model deployment represent only a fraction of the AI infrastructure's lifecycle. An AI system consists of eight major components: services, models, datasets, feature pipelines, feature stores, feature templates, models, inference services, and service orchestrator. The main components of an AI system are:

1) AI services: The supported online or real-time services provided by the system, such as anti-money laundering service, financial crime service, intelligent investment service, call center bots, virtual loan assistants, information retrieval service about financial events, and cybersecurity service.

2) AI models: This includes the AI model that generates features and the sub-models that support the model pipeline.

3) Datasets: These are collections of values or observations for either a single feature instance or for many instance features. The entities describe the features, contexts, or labels. Datasets are composed of features, composed of attributes. There are feature values that are composed of data examples.

11.3. Key Components of AI Tech Stacks

Financial institutions could develop AI tech stacks by unifying the key technologies and choosing an efficient process to manage large data transformations. In order to let AI contribute to the primary goals of the banks, a scalable AI infrastructure is critical for financial institutions. First, we recommend the AI tech stack in assessing the state of data availability and management, and discussing the primary data stacks, which support banks' customer strategy and data architecture. Next, we explore the future tech stacks by introducing emerging technologies that would be beneficial for fintech or banking sectors to speed up the process of high-level model iteration. Finally, we also introduce the techniques and cases utilized in complete financial data lakes to develop a user-friendly architecture and interactive analytics dashboards with highly scalable storage and computing capabilities, aiming to satisfy the requirements for developing professional AI models.

The common high-tech intelligent computing stack might involve machine learning/AI models, in-memory distributed NoSQL databases, and model deployment frameworks with cloud computing. It is worth mentioning that various layers need even more

intelligent middleware to break the silo effect in the bank. Transfer learning models would be the most efficient way since they reuse knowledge gained in one model and apply it to rare events seen in another model. It is a widely recognized challenge for financial evaluations. Model deployment is a framework to make it easy to engage data science on any task or customer profile's requirements. Ensuring that once the models have processed the data, they can learn faster next time when processing the information for the model. Once the models are developed by data science or AI, instead of sharing the technology between teams, the transfer learning method keeps the model reusable. With real-time model deployment, users can access and interact with the model as they see fit via this real-time model framework.

11.3.1. Data Management Systems

The requirement for real-time scalable data management systems is becoming more crucial as compute and storage layers inflate far beyond the petabyte mark. Generally speaking, the financial data-handling process is divided into a few steps, including data ingest, data quality controls, and data delivery. A data management framework must support any level of data movement efficiently while maintaining low control checks. With modern big data systems supporting SQL and indexed databases, lazy requester-issued batch data movement processes are becoming an antique subject. Simplicity in backend data access is a key feature in a data management system, so AI applications can get data quickly and efficiently with particular time-to-speed-read performance in mind. Relatively recently, graph databases have garnered quite some importance, mainly with synchronized distributed execution hardware designs based on fast increment and decrement operations. Unfortunately, neither specialist hardware implementations nor fast machine learning algorithms that work with these graph databases are available, so we focus on flat column-store implementations.

A backend layer capable of pushing down analytical queries or specific AI nodes embedded into the database layer for quicker calculations on ports has the appealing notion of guaranteed sub-second latencies in the future. Securities data handling requires specialized AI feature handling, but the principle works well here, as we can serve customer features as ARBars regardless of whether we get the security features from a time-series data manager or load them from rectangles. Data partitioning of tables to adjust to cache and index size limitations is standard. Ingesting is a more elaborate problem in a big data scenario, as we need to feed partitioned historical data, near real-time streams, and typical real-time AI requests with different latencies, aggregation levels, and accuracy. These can be resolved into subflows and pushed into specialized routers to reach different AI devices with the right AR architectural features.

11.3.2. Machine Learning Frameworks

There are more than a few deep learning frameworks available for AI. A machine learning framework was built to implement various machine learning algorithms efficiently. The data distribution engine makes various data types perfect for training and prediction. The model serving module supports CPU, GPU, and distributed computing, where the Global Interpreter Lock is finally removed. The model and data version control modules help improve the quality and stability of automated machine learning. The monitor and log modules are fundamental features to be accountable for the AI model in the regulated banking sector. All components are integrated with each other for effective interoperability and scalability.

Machine learning frameworks help to implement various machine learning algorithms efficiently. A machine learning framework was built to continuously improve the business for various needs, like model predictability, precision, and stability. A binary columnar representation, in-memory for reading, writing, and processing data more efficiently, is a data distribution engine. It provides analytical support for the most data types and schemas. The model serving module, user-configurable via a lightweight file, automatically allows for the best performance with multi-module mixed computing. It supports CPU, GPU, and distributed execution, which is achieved with a method that uses a thread to scale across CPU and processes to utilize multiple GPUs. The Global Interpreter Lock is finally removed. In addition, the system is based on a microservices architecture, further implementing end-to-end microservices, integrating and collaborating global machine learning models. All the components are going through quality improvement continuously and are enhanced with new features, following the same processes and standards as the current business solution.

11.3.3. Deployment Platforms

Today, the de facto technology for deploying AI models is via some form of software containers. These containers could be directly created using tools such as Python and R, or specific services. With the vast amount of AI models available today, these containers can be especially useful if focused on particular financial services. Businesses around each of these areas of model deployment are thriving today. With this trend of connecting technologies, their target is shifting from sophisticated data scientists and machine learning practitioners to more general business users. In the near future, businesses see a trend that tens of millions of people worldwide, especially those in the financial and healthcare sectors, will be able to train and deploy their own AI models with simple clicks on their web browsers.

However, we are also seeing the increasing complexity of AI models. As mentioned, many of today's models are created by AI models themselves, which can be very complex. The explosive growth of alternative data sources combined with rapid AI innovation is creating further pressure on financial institutions to adopt or be left behind. Regulatory, explainable AI, and model risk management present additional hurdles. Assembling complex models drawn from a vast array of different components, such as dynamic data analysis, sophisticated or custom machine learning or AI models, data selection, filtering, and hedging, will require an end-to-end deployment platform that does much more than simply wrapping these models into containers or adapting them to become data services. The platforms should tackle issues traditionally referred to as ModelOps, Model Retraining, Model Re-tuning, Model Updates, and Model Generation or Innovation Platforms.

11.4. Data Acquisition and Processing

Data, as an essential input to AI processes, fuels not only algorithmic decision-making but also improvements in the quality of the decision-making process itself. However, the massive increases in data volume and in the number of sources of pertinent external data, when combined with the regulatory and infrastructure constraints faced by financial institutions, complicate the data acquisition and data access processes. As a result, financial institutions find it a continuous challenge to both ingest ever larger amounts of externally sourced data as well as to create and maintain unified, comprehensive databases of data describing both the institutions and the customers of those institutions in a form that can be exploited for AI-supported decision-making.

The data acquisition challenge is in some ways quite simple. More data from more sources in more structured formats is better. However, in practice, regulatory and governance demands placed upon the data acquisition process have significant complicating implications. While a financial institution might very much want to acquire all sorts of wide ranges of data on a wide range of topics, that institution will frequently be forced to refrain from doing so or to at least take steps to anonymize the data before it is collected and later during its processing. That frequently implies that interaction with the providers of the desired data and the external data vendors must be involved. Furthermore, the challenge of data access is exacerbated by the ever larger costs associated with the process of acquiring, scrubbing, and combining data from multiple sources. Like data acquisition, data access is both expensive and complicated.

11.4.1. Data Sources in Finance

Finance, like other domains, faces unique challenges in terms of its AI infrastructure. Specifically, financial institutions have to deal with numerous proprietary data sources as well as unique regulations. AI in finance should always leverage as much internal and external proprietary and reference data as possible. The data landscape in finance is particularly fragmented, as there are numerous proprietary data sets, including customer information in customer relationship management systems, transaction data, and trading volume from market data vendors. AI use cases differ among financial institutions. Insurers may be most interested in understanding branding and customer churn, while banks have additional AI applications in anti-money laundering solutions. The amount and quality of available credit default swap data can vary by bank and by geography in the case of international banks. Hedge funds use AI for short-term alpha, including risk and event-driven strategies. Meanwhile, fixed income players use AI to extract signals from text news services. AI tools in these use cases leverage many data sets, including general text, financial news, companies, and additional reference data. Access to proprietary data can create competitive advantages. A data strategy, including third-party data usage policies and vendor negotiations, is important for ensuring that an organization has an accurate and holistic data set. In addition, it is important to obtain and use investable test data sets to have an accurate AI model.

Preprocessing and featurizing of these finance domain data sources are critical. Data dictionaries are helpful for complex and large data sets. In banking, the primary data sources are transactional data, system logs, and customer CRM data. While market data is often exogenous, fundamental data can come from reference data sources. Data can thus come in multiple formats. Event-driven data tends to be time-aligned, while textual data has dynamic time stamps. Fintech presents additional data management challenges due to real-time streaming access. Certain best practices can be used when working with finance domain data. Data in banking and finance is often hierarchical in nature, with accounting books as the primary nodes. Additional data can be added to facilitate efficient time travel and feature sharing mechanisms that incorporate alternative data. Machine learning models facing market data challenges due to tick size issues often need to normalize the data. Modeling challenges exist with accounting-based quarterly data. High-frequency alpha signals are important, but less attention is given to classical performance indicators. Classification models trained on quarterly data but scored on daily level may have label issues. Data set balance issues and the choice of performance measure can be complicated, while factor cycle usage can help in dealing with model performance issues. Post-launch, time-varying feature importance plots allow for data and model monitoring. Establishing a holistic audit trail is also important for regulatory purposes. During data and model build, both known and new data signals need to be appropriately labeled. In terms of what comes next, retrieve records and extraction logs are important, while scripts present issues and having data definitions for the compliance team’s understanding is important. Challenging data queries can create vulnerabilities to

bad behavior. Finally, it is necessary to normalize the feature set so that each model supports the same factors and maps the discrepancy.

11.4.2. Real-time Data Processing

Real-time data processing refers to the processing of data that includes operations on a record or document scale, with the ability to immediately render operational results to users or connected elements of systems. In contrast to batch processing, where data are gathered for a period and then processed, real-time processing renders results at a more granular level or with substantially reduced delays. In the context of financial services, real-time processing of data gives institutions the ability to respond to changes faster, to better manage risks, reduce fraud, optimize trading, and act on opportunities faster. Real-time analytics create the ability to aggregate data and perform queries across multiple users, sources, or formats almost instantaneously, leading to real-time actionable intelligence.

Real-time data processing is enabled through a specific system architecture or event architecture, and the principles form the fundamentals upon which successful real-time data processing systems are built. The basic pattern relies on collecting data, publishing event data, identifying events of interest, extracting relevant information, and reacting to the identified events. Systems are built in layers, with 'hot data' that are centrally managed forming the base layer, and 'cold data', which need some level of processing, stored on edge computing resources or batch processing systems forming the next two layers.

11.5. Machine Learning Models in Finance

In recent years, there has been growing interest in the use of machine learning approaches in the financial services industry. In the realm of credit lending, ML models are increasingly implemented by banks to streamline credit approval processes, improve the accuracy of credit risk assessment, and reduce potential losses related to bad loans. In retail banking, ML models are used to detect fraud in transaction processes, improve customer segmentation, and boost recommendation engines. Although the applications of ML techniques seem plentiful, they have been primarily implemented by top-tier banks, and several major issues prevent such deployment from being scalable to all financial institutions. These issues include the AI hype, lack of elite data scientists, resource constraints in utilizing ML architectures, and regulatory limitations imposed by emerging global regulators. If the majority of financial institutions face the same problems, each proposed ML solution by individual AI engineers will have insufficient practical appeal to achieve business scalability and attractive incentive schemes.

While ML models can create huge growth opportunities for the financial services industry, designing scalable infrastructure to support the ML model production pipeline and operation deployment within each process securely and transparently should not be overlooked. The real challenge is not ML model development, as frameworks currently lower the technical bar for data scientists. Yet, the promotion of model diving and deployment cannot be limited to access to elite data scientists. We provide scalable AI infrastructure that can help small to large-sized financial institutions solve end-to-end AI challenges within their business and operation processes. This enables collaboration with insurers to perform a holistic model performance evaluation with external production data and ensures that any potential issues are identified and resolved prior to deployment.

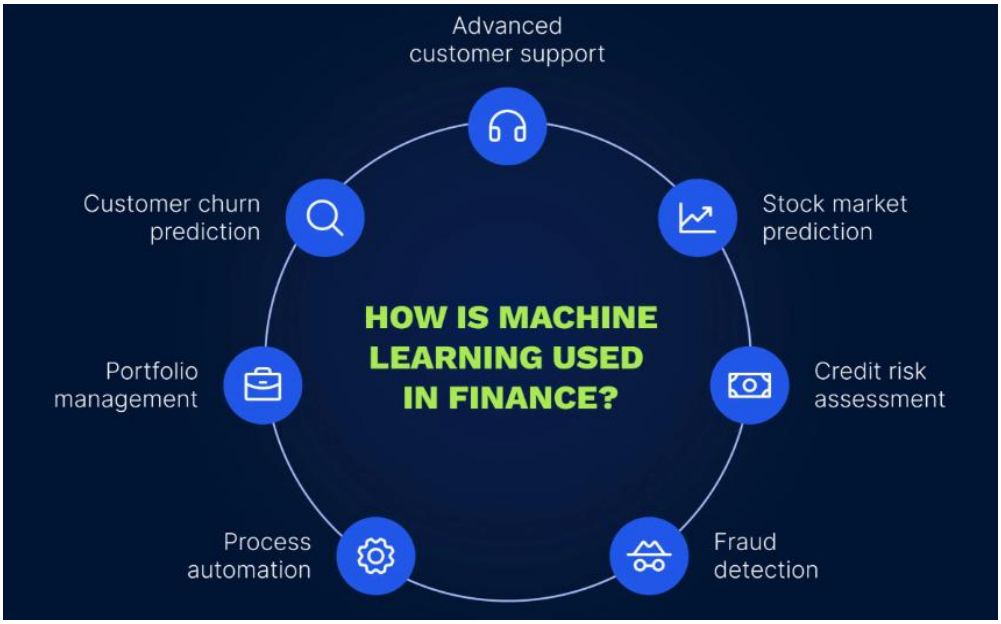

Fig 11 . 2 : Machine Learning in Finance

11.5.1. Predictive Analytics

Financial institutions have a large set of time series data and retail customer datasets on customer spending and usage behavior. We provide the capabilities to onboard, curate, prepare, and reduce the dimensionality of these datasets, in addition to building machine learning models and scoring against this massive data. There is built-in support for financial analysis and various generic and domain-specific ML and forecasting models. For example, users can leverage methods from time series, matrix factorization, and

association rules to preprocess their datasets and utilize ML algorithms to gain the required insights from their data. Users can compute the importance of the features in their dataset.

We provide capabilities to apply ML algorithms to financial data, spanning tasks such as time series forecasting, matrix factorization for co-occurrence analysis, and customer segmentation. Primitives to detect sequential patterns in data help with predictive maintenance and learning associations between different sequences of events. Such algorithms primarily work on transaction data and can be applied to financial usage patterns, helping with anti-money laundering and fraud detection tasks. Patterns of financial transactions corresponding to different types of events help organizations gain insight into the types of customers that are prone to certain types of activities. Such insights can be used for acquisitions, defining sales strategies, and improving customer experiences.

11.5.2. Risk Assessment Models

Risk assessment models are important quantitative models that evaluate the risk and return characteristics of investment positions. These models are also important inputs for portfolio optimization and risk/return attribution models. Typically, risk assessment models are built to assess the risk of a specific asset class or a specific type of investment instrument for a specific set of financial institutions and/or a specific set of investment strategies. In this section, we will discuss two classes of risk assessment models.

The first class of risk assessment models is the fundamental and statistical-based global investment and risk/return model. This model is built to factor returns of global assets, such as fixed income, equity markets, derivative indices, and other alternative investments, into common and intuitive elements: forecasts for fundamental macroeconomic state variables and investment style factors such as value, size, quality, momentum, growth, earnings, balance sheet, and technical factors. This class of risk models is typically used to assess the factor risks of an asset class and also to assess the risk profile of passive, strategic, and/or overlay investments using the specific factor exposures implied from global factor returns. These models are mostly macroeconomic models and use a straightforward cast of market-defined instruments such as futures, forwards, equity indices, country-specific government bonds, credit, inflation, and exchange rate forward markets for a distinct investment horizon of the financial institution's investment universe, with a particular emphasis on the rolling three-month investment horizon.

11.5.3. Fraud Detection Algorithms

Artificial Intelligence (A.I.) technologies can automatically learn and analyze data at a large scale, detecting latent associations and making predictions beyond human capacity. Adopted widely by various sectors, A.I. technologies have recently shown significant impacts within financial markets. One notable characteristic of the financial market is the necessity to perform extensive data processing and analysis in real time. Unlike technology giants and IT companies that could effortlessly deploy A.I. systems on cloud-based platforms, financial institutions often lack the capacity to efficiently scale A.I. systems due to inflexible data center infrastructures. However, confidential data and data regulations strictly limit the range and structure of A.I. facilities that can be deployed. To break this deadlock, we implemented A.I. infrastructures with a new scale-out architecture. We constructed A.I. PaaS and A.I. IaaS facilities that solved conventional performance limitations, providing financial institutions with extensive data processing and analysis capacities across all asset types and data horizons. Our simulations and case studies demonstrate the significant performance improvements and visible future advancements.

Financial crimes impose substantial costs on financial institutions and societies at large. The realm of fraud detection, which aims to identify criminal attempts in advance, has been a long-standing topic in the financial industry. Historical deep learning approaches, such as the recurrent neural network and regular feedforward multi-layer perceptron, have been proposed to detect potential fraud activities. However, these proposed solutions encountered low detection accuracy and posed strict sequence maintenance criteria. In particular, fraud detection is often presented as a binary classification problem, where features are aggregated based on low-level schemes such as windowing or aggregation. Theoretically, this measure means that the input sequence cannot be too long. Also, existing sequence learning architectures, such as recurrent neural networks, have limitations when sequences are too long. In this study, we address this measurement issue by using and comparing different types of architectures. Our experimental results provide insightful performance precision trade-offs for different fraud detection tasks.

11.6. Infrastructure Design Principles

In this chapter, we detail the design principles for the scalable AI infrastructure that supports the use of machine learning by many developers at JPMorgan Chase. Few such networks exist, and the size, scale, or security sensitivity of data that can be processed dwarfs the challenges at most other companies. These principles have been established over a period of constant learning from a multi-year process of setting up a clear centralized hosting infrastructure for ML in the bank. The key insights apply to many

infrastructure engineering settings, but the levels of security often involved might be new to many sections of the machine learning community.

The principles involve identifying the true cost and true value of all ML and using these principles to construct a space offering at their intersection that meets the business and users' needs. Also articulated is a culture of proactive collaboration and high-touch engagement between ML infrastructure engineering and internal service customers. To illustrate and explain the principles, the design of a large machine learning infrastructure is discussed, detailing a viable use case from the consumer banking segment, and providing both data and compute infrastructure support for training at scale. These principles are important because scalable ML infrastructure is more than just a cluster going from single digits to triple digits of GPUs. It is about ensuring that a rapidly increasing segment of the business can trust that the tools they are choosing to use are supported by a stable, rigorous, and secure platform.

11.6.1. Microservices Architecture

A microservices architecture promotes implementing complex systems as a suite of small, independent services. These services run as distributed services and maintain a focus on specific business functionality. Each service is fully decoupled and runs its own processes. A microservices approach achieves a few interesting properties that help manage a complex system. First, by partitioning the system into loosely coupled services, we reduce both the code base and the complexity of the project concerning both development and testing. This leads to increased agility in both feature release and iteration. Second, microservices can be built with their own language and technology stack, making them more flexible to deploy. This not only simplifies deployment but also provides greater durability to technical design stack changes to the team. Finally, microservices scale to handle load more effectively. Every service is built to scale, and the integration of all different services together provides a highly scalable ecosystem.

An architecture like microservices offers a number of benefits, but in the AI world, microservices provide two additional advantages. First, microservices provide a natural abstraction for sharing models across different products. The domain knowledge encoded in the AI models could be deployed in different microservices. For example, a recommendation model for use in both your site and in your chat system can be deployed as a microservice and reused by many AI-supported services. Second, popular deep learning frameworks all integrate well with an open-source model serving system that provides a microservices structure to the system by wrapping the AI models in very lightweight containers. By using microservices, there is the flexibility of mixing and matching models built from different AI technologies being deployed in the same environment.

11.6.2. Containerization and Orchestration

Docker is a famous tool to manage containers. It is designed for developers to build fast and light containers leveraging the Linux kernel. After creating an image, it can be run in a container with isolation. This allows a developer to be sure that their application will be correctly executed on any platform like development, testing, delivery, or production. Docker also provides a registry that allows distributing the images. Kubernetes is an open-source platform designed for large distributed systems. It uses proximity semantics based on container workloads.

Financial institutions will use technologies such as Docker to build their own private cloud to run applications in their own data center leveraging their existing infrastructure. Kubernetes will be used to deploy, scale, and auto-manage these containers. With AI techniques in the infrastructure field, it will be possible to execute intelligent data processes over all data in the company. In this context, financial institutions should also provide regulated and secure pipelines for data scientists. Results of these models can manage AI models, leveraging a complete architecture already deployed in-house.

11.7. Cloud vs On-Premises Solutions

Financial institutions have always had the choice between vertical versus horizontal scaling, and owning versus renting their infrastructure. On-premises solutions are custom-built, capital-intensive, and expensive to deploy and maintain. Cloud solutions are off-the-shelf, OPEX-intensive, easy to deploy, and equally expensive to maintain. Financial institutions rely on classifications and reports to guide their decisions between these extremes. The right balance for each institution depends on its business demands, bottom lines, strategic focus, and IT capability.

The biggest factor in favor of leveraging cloud-based technologies is the ease of getting started. It is deceptively easy to experiment on the cloud using modest budgets, developers, and IT staff. Much of our recent innovation can be attributed to the rise of cloud-based solutions. Organizations can now subscribe to services that they were unable to develop, deploy, or secure by themselves. The comparison between the costs of cloud computing versus existing costs of on-premises solutions is striking. Given the shrinking race for zero day to connected days, many institutions now recognize that considerable expenses on security research and countermeasures are ultimately a good deal for them. Head-to-head feature comparisons with on-premises products are beyond the scope. The only larger head-to-head events involve startups developing scalable AI-driven solutions from scratch versus incumbents with large tech budgets and middling IT capabilities.

11.7.1. Benefits of Cloud Computing

The previous section focused on the benefits and challenges of developing in-house AI/ML infrastructure. In this section, we compare this approach with leveraging cloud computing infrastructure management by paying a third-party service and its implications. The main benefits of cloud computing are cost reduction, scalable performance, and time savings. Cloud platforms provide on-demand computing resources to run AI/ML models and storage for data to train and predict AI/ML models. They also offer the flexibility of paying only for what a user consumes.

Another benefit of cloud computing is the ability to instantiate and terminate parallel processes. For AI/ML models to train, their scalable behavior is important to reduce the time to obtain business insights and decision-making benefits. Cloud computing platforms can instantiate computing resource outputs on demand and terminate them after the model is trained. This is particularly useful for tasks that need to be performed quickly, resulting in a favorable cost/time relationship. Moreover, parallelized computing setups in which inexpensive and massive numbers of programming threads are integrated with the cloud offering can result in excellent performance improvements right out of the box to train AI/ML models without the need for complex programming overrides.

11.7.2. Challenges of On-Premises Infrastructure

We asked several financial service institutions about their approaches to managing their AI infrastructure and operations. One large bank has set up a multi-cloud platform with their IT resources running on-premises in a data center, and for backup and for legacy use case support. As part of their challenges in running their AI on-premises, the bank listed compliance, networking, on-premises system capacity, and scaling for deep learning training. They have to keep AI data on-premises, as the preferable combination of consumer protection record data with sensitive AI predictions makes it hard to get these datasets out of their system. Consumer personal information protection regulations and banking data protection regulations mandate that the bank maintains compliance with special controls, including a security baseline and user access audits handled in the data container to ensure that criminals do not take advantage of external threats.

However, the on-premises networking device used by the bank has very limited bandwidth, and there is no AI-specific compute device available. Although they also take advantage of GPU and edge devices, the network, including the two on-premises networking devices, plays a leading role for their prediction devices. After reviewing the networking performance and the above limitations, GPU setup, and considering that many prediction services will run on a micro-batch of the GPU due to using a way, the

bank decides to reorganize each GPU into a set of GPUs since inter-GPU communication is not the focus of their banking workloads. Family groups to overcome limitations. The group setup has both CPU and GPU support and satisfies the high volume capacity required by inference services. To support both deep learning and routine workloads, the larger servers support both running on a Kubernetes environment and non-Kubernetes labor. For edge devices, a combination of personnel from the bank and integration projects from cloud service providers can implement the on-premises solution. The bank operates its AI services on either its three types of on-premises infrastructure or on. Moving workloads creates the prognosis capacity that they would otherwise not get on-premises, and as their initial training device, is a solution for testing defectors, but most live predictions cannot be completed in the cloud. Their AI strategy will provide nimble multiple projects, balancing the overhead of the cloud. AI capacity purchases on-premises and skills to compensate for the bank's network limitations will let the temporary business benefit from gradual onset instructions. To address the new challenges on the roadmap, the bank decided to hire architectural skills with business development capabilities and specialized knowledge in on-premises and edge devices, full stack communication, and data pipelines, ensuring the data's AI model validity and security during its lifecycle.

11.8. Security Considerations

To ensure solutions built with BigDL maintain the security of sensitive data, such as investment strategies, BigDL leverages secure communication algorithms provided by both MKL-DNN and open source projects. When running deep learning frameworks in production, extra security measures must be taken to protect deep learning workloads, training data, and models, or else it will be easy for attackers to leak sensitive data or tamper with models. Full stack security considerations are necessary to protect all layers, including both the underlying infrastructure layer and the deep learning specific algorithm layer.

First, isolation between workloads is a fundamental necessity that is addressed by cloud providers; these enterprises provide a multitude of platforms to allocate independent space to different business organizations or architectures. Otherwise, vulnerable deep learning infrastructures would open the door for a broad range of post-exploitation scenarios. Enterprises that have their own private deep learning infrastructures should build isolation mechanisms similar to those employed by cloud providers. Second, secure communication between different components in the cloud and deep learning frameworks is also critical. End-to-end encryption and cryptographic methods used in data security schemes are leveraged to establish secure channels between deep learning frameworks and services interacting with input data. With an emphasis on big data and

deep learning infrastructure, this text articulates security challenges that need to be recognized and tackled thoroughly by focusing on cloud and on-premises deep learning platforms. Guided by the principles of security by design, our exploration incorporates the elaboration of specific threats and the frame of actions designed to eliminate them in both cloud and private infrastructures. Although the text provides a comprehensive exploration, it focuses more on the cloud environment because of its popularity among deep learning projects. Nowadays, many organizations leverage the cloud's cost advantages to train AI in a serverless manner.

11.8.1. Data Privacy Regulations

Scalable AI infrastructure for financial institutions is a challenging stance due to the continuous evolution of data privacy regulations. The constantly evolving and stringent data privacy regulations require financial institutions to process and interact with the data in a compliant manner. These regulations directly influence the technical infrastructure decisions. Non-compliance with these regulations results in regulatory violations and monetary penalties; at the same time, financial institutions cannot compromise on the developments and advancements in the artificial intelligence vicinity. Major financial institutions face these challenges.

Legal representatives have analytical expertise and keep the company up to date with data privacy regulations. They propose control measures and tools to adhere to these laws. These laws are predominantly named as standards, and companies that are compliant could also consider themselves compliant with other data privacy laws. This interpretation may work at a high level but is not correct if we look at the implementation and penalties. The data privacy laws are becoming stricter, and business processes, tools, and standards developed for one data privacy law should be revisited before assuming compliance with the newer laws.

11.8.2. Cybersecurity Measures

This assumes that each infrastructure handled by financial institutions is vulnerable to repeated cyber attacks. Specifically, the infrastructure supporting AI services is likely to become an important target of attacks due to its core position in terms of aiding the decision-making process. The emergence of AI-specific security threat vectors is undeniable, and malware threats targeting AI infrastructure are in an accelerating growth phase. If AI models and training data are crucial assets, it is natural that these assets must be adequately protected to enhance a competitive advantage in the sector. This means that banks must build and maintain a secure AI ecosystem that protects all of their AI assets.

In line with this trend, a growing number of factors under Security-Sensitive AI Market include AI Infrastructure Security, AI Model Poisoning Attacks Defenses, State-Sponsored AI-Enabled DDoS Attacks, Adversarial Attacks on Deep Learning Machines, Adversarial Attacks at Black Box Payment Fraud Detection and Mitigation, Explanations of AI Systems. With the success of adversarial attacks at evading real-time detection systems in multiple AI-based industries, exposure to hackers and criminals will continue to increase. It is hard to completely prevent malicious actors from exploiting the data and models of AI. However, the financial industry, in particular, is in danger if the data is biased, corrupted, or manipulated. The fact that some organizations misuse their AI-powered technology is not an unknown issue. If it is known, then the basic question of why it is allowed to do so must be asked. Then there is the consideration of what will ensure the effectiveness and protection of AI. In AI, all the way from initial training data collection and model development to ongoing operation and adaptation, there are many places to inject errors, biases, adversities, and vulnerabilities. Given the scale, scope, and the number of decisions AI now supports, any kind of attack can have enormous consequences.

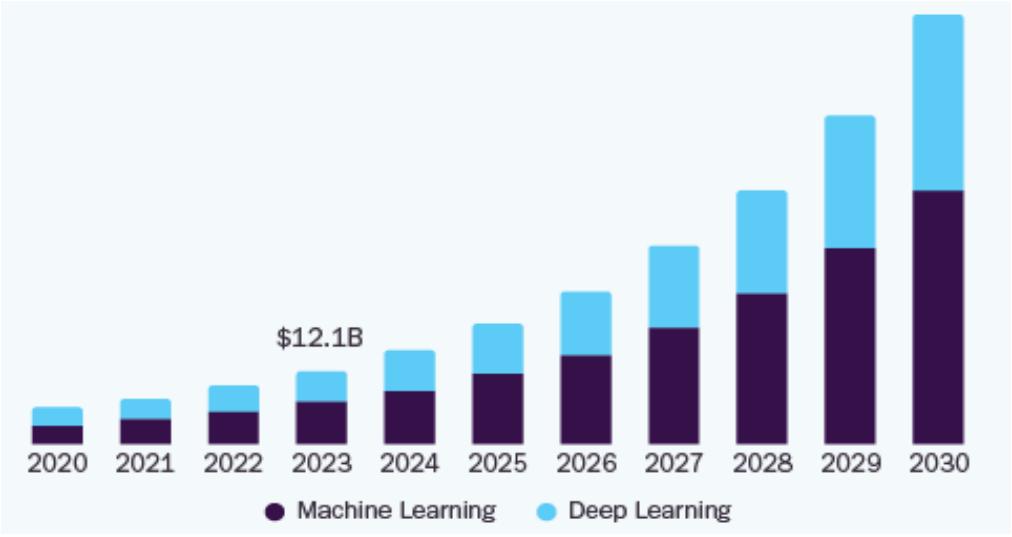

Fig 11 . 3 : AI Infrastructure Market Size

11.9. Integration with Legacy Systems

While it can be argued that the integration layer should be able to adapt to any technology of the existing system, currently only those integration problems that are common are being tackled first. From our extensive list of known problems, these technologies should then be enhanced to cope with the unforeseeable part of the legacy issue. Open standards would open many possibilities for the development of such systems. The major

challenges in dealing with integration are issues of data independence and remote processing. These imply that the architect building an integration system is constrained by the rapid changes occurring within all systems and needs to develop localized solutions with limited vision before embarking on a full-scale development plan. While we can define a general minimum set of interfaces or functions that are common to all systems, they have to be small module building blocks. The diverse architectures of systems can be carefully controlled. Key to the success of a new architecture is not to assume future development of the systems that these modules will connect to.

Thus, a standard interface library that is separate from the actual application design is required. Let the engineers worry about the detailed aspects of integration. The problems associated with each of these issues can therefore be isolated to one level. However, this standard set of integration libraries is simply suboptimal because performance gains would be lost at every level of the development by chasing down every last feature and every last exception to the rule. As such, efforts should be concentrated on working with the integrator and building quality interfaces, which go beyond the minimal set required to complete a given contract. If systems ever reach that stage of perfection and identity, customers and companies would become interchangeable, and the question would become less about how to choose among equally excellent companies than how to get the cheapest, or perhaps just how to get the legal one to fulfill the contract.

11.9.1. Challenges in Integration

The vision for integrating AI with business operations requires overcoming a number of challenges, including management of large-scale AI infrastructure required to train and deploy machine learning models. The AI infrastructure problems are exacerbated in financial services, where there are typical problems with scaling deep learning, such as model customization at scale, tuning and training many constantly evolving models, and the ability to deploy optimized models, coupled with additional challenges due to strict data governance requirements and the need to deploy on a number of interconnected legacy systems while meeting strict security requirements. Many data scientists and engineers are required to manage these pipelines. In the end, the model might only provide a minor uplift over more traditional machine learning models.

Data presents similar challenges. Banking data is complex, highly interconnected, and contains many sparse features. Models need to calculate quantities such as customer lifetime value, which is typically a large aggregation of many types of events, such as a deposit, a withdrawal, or a click on a website. However, traditional data protocols make it difficult to calculate these quantities in an operational time frame and store the results at an atomic precision required for model training. The result is an extremely long

iteration cycle for feature creation, and often traditional algorithmic solutions end up being more effective for production models.

11.9.2. Strategies for Successful Integration

As financial institutions continue to seek new ways in which AI can create value for the business and enhance their market position, successful integration of AI techniques will only become more important. By recognizing the potential challenges associated with AI integration early and by addressing these challenges through strategic planning, we believe that these institutions are in a better position to make use of the wealth of techniques available to them. The methodologies and strategies presented promote a clearer understanding of the nature of AI techniques and how these techniques are actually integrated within a financial context. By thinking and working with AI in this way, more practical and effective use of AI techniques will result.

AI techniques have certainly changed significantly over the years, shifting from deterministic rule-based approaches to more statistically driven techniques such as data mining and machine learning. AI integration has similarly changed, driven by advances in framework design and by developments in computing power. Successful integration of AI can leverage these techniques as a necessary foundation. However, we also suggest that truly successful AI integration will require an evolution in the capabilities required for integration – an increased need for an innovative approach toward the integration of AI with business processes. We conclude with the suggestion that the most successful practitioners of AI in financial contexts may well be those that see AI as a technique that can make a broad range of computational functions more effective and efficient, rather than seeing AI as having only a limited set of specialized applications.

11.10. Conclusion

This paper discusses two important pillars involved in building a large-scale AI infrastructure in financial institutions. They are: i. Managing AI workloads in an existing private/hybrid cloud environment where applications within financial institution operate, and ii. Building performance and cost-effective machine learning solutions over big data prevalent in financial institutions. Financial Institutions across the globe undergo a digital transformation to innovate and grow, especially by focusing on customer experience, streamline processes and operational effectiveness, and take smarter decisions. This interaction with clients and ecosystem partners results in huge data. Besides, capital markets organizations have made significant technology investments in databases, distributed data processing engines, cloud computing, and more.

These technology advancements resulted in financial institutions having Big Data, AI, and Cloud mandates as some of their top technology themes. These are driven by business imperatives such as reducing time to deliver projects, better cost-effective ways to test new ideas, deliver innovative experience, and develop better niche operational advantages vis-a-vis peers. We provided major decision criteria and recommendations for augmenting an HPC or, more generally, a cloud infrastructure with GPUs for DL, and described after the fact tunings. FinSimML outperforms the related CPU version. Also, a reduced precision approach delivers competitive results, both at training and inference stages. Therefore, it is an alternative to implement models in real-time scenarios with more demanding system constraints. Upcoming work includes adding a regression model for fine-tuning results, considering the λ values as hyperparameters. Finally, FinSimML can attend to technology demands or curiosities concerning this ever-growing domain.

11.10.1. Final Thoughts on AI Implementation in Finance

The ongoing implementation of AI in finance, based on our own and our customers' experiences, has explicitly motivated a new axis for taking our scalable AI to the next level. Still, various new challenges are causing peculiar problems in legally regulated sectors such as finance. We have pinpointed the particularities related to the actual industrial implementations of AI in finance and proposed our scalable industrial AI equipment for practitioners to meet such challenges. As a result, we have also taken another step towards a more successful AI-driven future. With significant challenges remaining, the financial services industry may continue to play a leading role in AI implementation, this time focusing on more specific challenges and opportunities in AI research. We hope that our AI equipment may contribute to improving AI technology, reaching complementary machine learning and other natural language processing. AI for financial services aims to enhance actionability and spread the benefits of AI to a broad enterprise or industry. The problems discussed are absolutely timely. Our AI equipment is not the final answer. In the evolving landscape of AI and finance, our scalable AI has just taken on a new dimension.

References

Hameed, I., & Yang, C. (2020). Building Scalable AI Infrastructure for Financial Institutions: Challenges and Solutions. International Journal of Computer Science and Information Security, 18(5), 118–129. https://doi.org/10.1007/s41036-020-00219-6

Ghosh, S., & Albrecht, H. (2020). Artificial Intelligence in Financial Services: Opportunities and Challenges for Building Scalable Infrastructure. Journal of Financial Technology, 3(1), 1–14. https://doi.org/10.1016/j.fintech.2020.02.003

Aversa, E., Lamberti, F., & Cecere, G. (2018). Scalable AI Solutions for Financial Institutions: A Case Study in the Context of Credit Risk. Journal of Computational Finance, 22(4), 63–90. https://doi.org/10.2139/ssrn.3284759

Kshetri, N. (2017). 1 Blockchain and Artificial Intelligence in Financial Institutions. Journal of Business Research, 87, 216–227. https://doi.org/10.1016/j.jbusres.2017.04.008

Zhang, Z., & Li, T. (2019). AI and Machine Learning: Leveraging Scalable Infrastructure for Financial Data Analytics. IEEE Access, 7, 108645–108656. https://doi.org/10.1109/ACCESS.2019.2935789