Deep Science Publishing
https://doi.org/10.70593/978-93-49910-19-5_2

DeepScience
Open Access Books

Chapter 2: Building a scalable and resilient data infrastructure to power intelligent financial services

2.1. Introduction

We live in a world of data, one whose volume is growing at an exponential rate. Recent estimates indicate that the volume of data created globally doubles approximately every two years, an increase driven in equal parts by the growing digitization of all sectors of the economy and by increasing improvements in the speed, resolution, and functionality of digital recording instruments. Financial services and technology firms have recently constructed large-scale data infrastructures to process the data from this explosion, in efforts to lower costs, improve revenues, and mitigate against risks (Cilliers & de Waal, 2017; Inoue & Matsubara, 2018; Chan & Mukherjee, 2021).

These data infrastructures collect, store, analyze, and visualize data from different sources, in different formats, and of different types, including structured, schemaless, and unstructured data. The different sources include the financial or economic activity of firms, households, institutions, and the government; the thousands of financial and macroeconomic variables reported by various governments and multilateral institutions; the millions of financial transactions recorded every day around the world; the millions of messages, reviews and information published online by users on social media and collaborative platforms; and the news articles published daily by press agencies. The data infrastructures analyze these data to produce alerts on the changes in the states of the firms, customers, or institutions; visualize dashboards to monitor the key variables and trends; and compound signals to be used in risk models or forecasts. Insights from these insights help decision-makers in financial services firms and corporations, such as chief executive officers, chief financial officers, chief risk officers, and trading desks, make complex strategic decisions daily. As with any other technology, the adoption,

integration, and deployment of these data infrastructures may be governed by technological, economic, behavioral, financial, institutional, political, or regulatory factors (Pfohl & Gomm, 2019; Kowalski & Rungta, 2020).

2.1.1. Overview of the Study

With recent technological advancements, financial services companies are relying more on data to generate insights, improve customer engagement, and ensure successful competitive positioning. This has led to an increase in demand for scalable and resilient data infrastructure, which can provide timely action-oriented information to reap benefits. In this study, we provide a comprehensive overview of the current state of data infrastructure at financial services companies and review the key components required to build a scalable and resilient data infrastructure. We also touch upon the evolution of the data infrastructure in financial services and underlying drivers. And finally, we highlight the importance of adopting the right technologies to fully take advantage of the new data environment.

Financial services companies invest billions of dollars in building technology capabilities to deliver value to stakeholders. With recent technological advances in artificial intelligence, machine learning, and data management systems, financial services organizations are relying on data to generate timely insights to support decision-making. There is a growing expectation that data be increasingly available for analytical purposes to help facilitate better decision-making, which in turn generates value, increases revenues, and lowers costs. Over the past few years, organizations have embraced this data-led approach to varying degrees and invested heavily in their data infrastructure and technology capabilities. Data availability on its own does not guarantee a competitive advantage for financial services firms. Making business decisions based on that data, and acting quickly on the insights is what matters. Nonetheless, organizations have leaned into data as a core competency and driving force to enhance operational efficiencies and create value for stakeholders.

2.2. Understanding Data Infrastructure

Definition and Importance

Data infrastructure is everything that enables organizations to process their data and extract value from it. At its core, it consists of computing infrastructure – data centers and servers that perform the work – and storage infrastructure – disks and object storage systems where the data is kept. While compute and storage infrastructure is relatively straightforward, data infrastructure is much more than that. It includes software

components that make storage and compute infrastructure scalable and allocation which is efficient and dependable. Think about it – we've all sent an email before, yet behind the scenes that email is sent to thousands of data centers; and with reliability that's been perfected over decades even for distributed systems. Data infrastructure is also the systems, software, and tools that make dealing with and processing data easy and finding and understanding data easy and dependable. Think about how difficult it is to set up, validate, and test a complex streaming pipeline that ingests and transforms data from multiple locations. The broader data infrastructure is the set of tools, systems, and software to make stream ingestion easy to use and fail-safe. Or think about how difficult it is to configure software and systems to train ML models on hundreds of terabytes of data. Well-paved data infrastructure has the software, practices, and templates to allow data scientists and engineers to put their data models and pipelines in production seamlessly and ensure they give back predictions or classification results that are reliable.

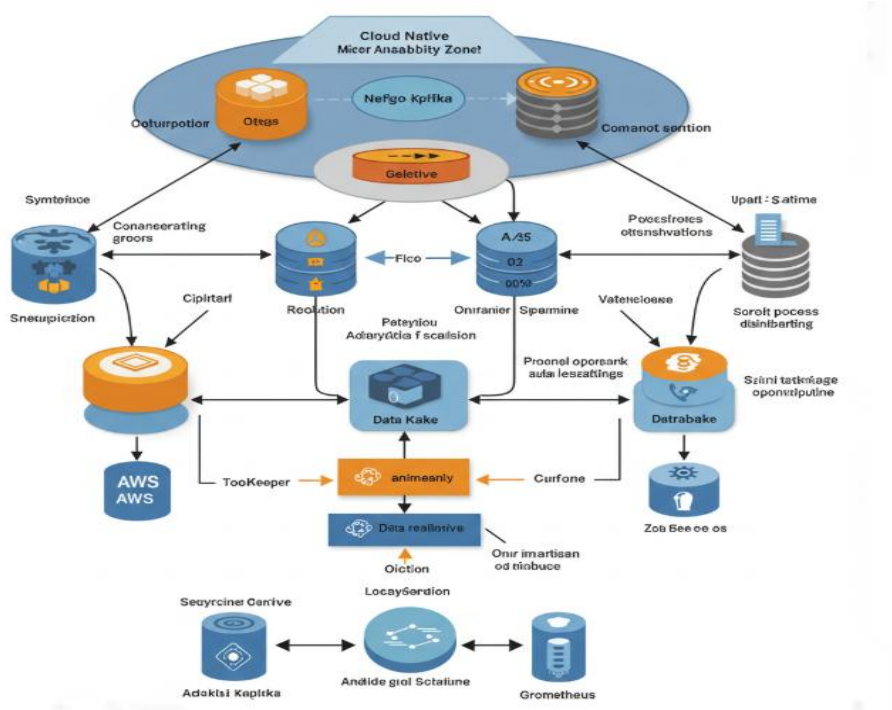

Fig 2 . 1 : Building a Scalable and Resilient Data Infrastructure

Finance is a space that's essentially a collection of decisions based on historical data. Since capital is priced according to time value and risks, those data are treasured. Financial companies, big or small, all transact high volume. Financial transactions are recorded on logs – of payments from one individual's bank account to another's; of trades bought or sold by an individual; of businesses borrowing money; of loans being

disbursed and payments being made throughout their tenor. Because of the nature of the industry, it necessarily attracts a large number of businesses and services that build products and add features on top of the core business – ranging from high-frequency trading, stock exchanges, risk and derivatives markets, investment banking, and trading, to hedge funds; and their agents – brokers who connect clients with investment companies. Therefore, with the growth of advanced technology and the move to digital, data infrastructure is not only a finance niche; it's core to the business.

2.2.1. Definition and Importance

Data infrastructure refers to the architecture and organizational components that interact together to collect, organize, process, and distribute data across systems and users in a cohesive way. Just as a physical infrastructure sets the foundation for transportation, housing, and communication within a physical space, a data infrastructure is a fundamental building block for the capabilities of a data-driven organization. The data infrastructure is a deep concept, composed of the key design decisions and constituent elements together that enable an organization to be data-informed or data-driven. At a minimum, a data infrastructure identifies and enables the fundamental capabilities of moving data through its journey in an organization, e.g., collection from disparate sources, transfer over networks, storage in databases, processing via pipelines and models, and distribution for use by stakeholders. As such, it is natural that data infrastructure has become a captivating topic, increasingly and quickly moving to the forefront of data and analytics organizations, not just in its management but in its large-scale design and architecture as well. While there are many articulated analyses of the state, trends, and challenges of data infrastructure, we emphasize that data infrastructure is essentially an interdisciplinary field, requiring input from engineers and operators, developers and users, technologists and business leaders alike to both design the constituent elements and understand and achieve the goals of infrastructure in a given organization. Data infrastructure enables an organization's vision for the acquisition, preparation, management, and analytics of data, to be realized consistently, seamlessly, and reliably at scale across many projects, domains, and stakeholders, and accordingly be a competitive differentiator and a key delivery pathway for organizational-wide capability enhancement and evolution.

2.2.2. Components of Data Infrastructure

Data infrastructure is composed of many critically important components that our enterprise architecture must accommodate. Unfortunately, many companies deploy data functionality piecemeal and without a long-term vision of a coherent data infrastructure

or architecture. They build operational data stores, data warehouses, data marts, analytics reports, streaming systems, enterprise platforms, cloud infrastructures, application program interfaces, etc. without regard for how these systems fit together or their performance characteristics. This can create data infrastructures that are more costly to develop and operate, unreliable, insecure, and difficult to extend with new functionality as business conditions or data needs change.

Your data infrastructure supports multiple functional areas. First of all, critical to any enterprise, and especially in regulated industries is the data security domain, which deals with providing accountability of data and data changes, including data encryption, recovery from corruption, and restricted access to only authorized parties with the ability to assess the security exposes and credentials necessary to access data in its various forms. Furthermore, certain critical business processes require enterprise-wide access to data that are cross-organizational in scope, making it necessary to design an enterprise data platform that allows supporting analytics and reporting functions that democratize insight discovery and allow business partners to work collaboratively with data scientists or data engineers in research and development of new capabilities. It would also require a cloud infrastructure with hybrid or multi-cloud capabilities to augment on-premise data processing in the case of temporary data-explosion conditions like monthly or quarterly reports.

2.3. Challenges in Financial Data Management

Servicing its customers promptly is fundamental to any bank's success. As needs arise, banks must provide customers with the financial instruments required to meet their demands. In doing so, banks create data mountains detailing the creditworthiness of all of their customers based on the history of their dealings with that bank and other banks. The customers demand that banks manage these mountains of data to ensure that their information is not subject to security breaches and that banks are compliant with the strict laws governing data management and privacy imposed by the regulations of the countries in which they are doing business, at the risk of experiencing significant fines. Inadequate security measures, resulting in the compromise of personally identifiable information, lead to distrust by the customer and consequently to the possibility of banks missing opportunities to create financial offerings that would benefit both parties. Although all businesses have their data management challenges to meet in the normal course of operations, financial services are unique in that they create a huge volume of extremely sensitive data that must be faultlessly secured the analysis of which could create additional revenue streams for financial services companies.

Data Volume and Variety

The problem you face as a data manager in a financial services company is that the size and diversity of the banks' data are growing far faster than your company's internal data management capabilities. To maximize revenue and meet internal reporting requirements while at the same time complying with external regulatory guidelines, data managers must deal with decision science projects that require a data infrastructure that could support four types of demands, each increasing in intensity as the volume and variety of banking data continues to grow – for production data updates, for exploratory data analytics, for operational reporting, and business intelligence activities.

2.3.1. Data Volume and Variety

Data volumes are also growing at a rapid pace. Though the term "big data" seems to be not so popular anymore, it is an apparent reality for many fintechs and traditional financial institutions. Most of our clients and our team ever work with terabytes of data on a daily basis. Data dementia is a widespread problem with financial organizations. The way we transfer financial transactions from one party to another creates tremendous amounts of data.

Financial services have become so diversified that any large player is dealing with many different types of products. The more diverse the portfolio, the more complex the financial architecture. For instance, such a complex corporation has multiple transaction systems, ERPs, Marketing Automation tools, CRMs, Trade Management Systems, Trading Platforms, Clearing Houses, CMSs, Data Warehouses, and Business Intelligence tools. These systems have been established over a long period and are specific to the organization. They have grown organically and been bought, merged, or developed by teams with different expertise, different priorities, different structures, and different flavors in different regions, lines of service, and capacities. Integration projects are expensive and take considerable time to run. They are often undermined by data variety and the number of systems, particularly at multinational banks and financial services institutions. This results in the logic diversity challenge, the variety of data models, business processes, governance structures, taxonomies, and metadata links. Poor data can cause regulatory compliance risks and advantages for business units that monitor and use data quality.

2.3.2. Data Security and Compliance

Financial data is probably the most sensitive data in all of the enterprise space. Security requirements put forward are some of the toughest. Financial institutions cannot afford

to expose data to attackers creating risks or misuse. Also, financial service institutions are highly regulated. Data privacy laws are very strict. All data must be managed in compliance with policies such as data protection regulations. These regulations help secure personally identifiable information including health data, financial information, etc.

Even when it is necessary to ensure security, information must still be available for analysis since it has business value. Due to the critical importance of data security, these use cases necessitate stricter data security, which includes rules and regulations on data storage and data movements, which may be at a finer granularity than secure data currently supported in data management tools or systems. So, the challenge is how to engineer an enterprise data infrastructure at scale, which builds safeguards against both internal and external threats, follows regulations, and provides an easy and quick way for users to prepare and analyze data that is still compliant.

Security is done at multiple layers in the data pipeline. This includes the data source, storage, processing, access, and governance. Rules must also be defined to provide a management framework where anything is either explicitly allowed, or not allowed, avoiding miscommunication. The foundation for security and compliance starts by classifying every data asset, along multiple metadata dimensions, including file types, storage location tagging, and tagging of sensitive content and data sensitivity. Filtering of sensitive data must also be done during data collection, ingestion, or extraction.

2.4. Scalability in Data Systems

The ability to scale data infrastructure to meet the growing business needs is something at the forefront of a technology leader's mind, perhaps even more than resilience and redundancy. Data infrastructures sometimes scale vertically by replacing existing data infrastructure components with machines that have greater capacity. This is often an expensive process. Virtualization has removed several of the cost barriers associated with major database or computing server upgrades. However, the cost bottleneck remains. In addition, for distributed computing systems such as big data processing systems or streaming query engines, vertical upgrades are limited by the throughput capacity of one machine.

Many organizations are adopting solutions enabled by public or private cloud infrastructure wherein systems can scale horizontally by adding additional commodity components. These operations—adding machines or instances rather than replacing existing machines—are software-defined and largely coordinated by external orchestration systems. As a result, virtual images of existing servers can be duplicated on demand to provide excess capacity for sudden surges in processing or storage

requirements. Also, configuring instances to work together as one logical unit provides horizontal scaling support for workloads that exceed any individual node capacity.

Solutions built on horizontal scaling have a lower barrier to entry than those that require vertical scaling approaches. Industry analysts are predicting that spending within this horizontal scaling space will grow to exceed anything deployed using vertical scaling approaches. In keeping with this trend, the introduction of cloud-based solutions on top of massive-scale existing infrastructure is providing a new reference point for pricing data systems. Organizations can start utilizing space and resources with very little or no capital investment. The outsourcing pricing model is the first option considered for a new project. These options raise the bar on the competitive pressure felt by product and service firms and internal infrastructure teams.

2.4.1. Horizontal vs Vertical Scaling

1. Scalability in Data Systems

Data is generated by financial and financial-related services at an unprecedented scale. Hundreds of millions of trading accounts generate tens of thousands of messages each day. These messages need to be processed, in real-time, to produce relevant and timely insights for hundreds of millions of consumers. The generated insights need to be prepared for immediate presentation in mobile applications or websites for users wanting to receive information on their financial assets. Additionally, reports on trading activities need to be made available and made easily accessible, electronically, to consumers at tax time.

This enormous volume of data can be dealt with only by scalable and resilient data systems that expand their capacity and withstand failures, just like the architectures of the microservices on which they are built. The scalability of data systems is linked to the number of requests that can be processed for a given volume of data. This leads to two primary dimensions of scalability—vertical scaling, which refers to processing requests with a single instance of a data system, and horizontal scalability, which refers to an increase in throughput by deploying additional copies of the data system behind a load balancer. Services that perform data management tasks, such as databases, are either vertically scalable or horizontally scalable. Some horizontal databases also possess limited vertical scalability. This paper will refer to the former types of databases as "traditional databases" and the latter as "Next-generation databases".

2.4.2. Cloud Solutions for Scalability

To understand how cloud solutions enable scalability without the need to invest heavily in new hardware, we must first at a high level see how they work. The cloud consists of several large data centers, typically with hundreds or thousands of machines, all tied together with high bandwidth and low latency networking. These data centers and the machines inside them can be leased in relatively smaller units or building blocks, depending on the provider.

The key to scalability with cloud service providers is that the networking speeds between machines in a data center and between data centers are high enough that one can combine machines in a data center to scale up services. At the same time, the inherent redundancy in the service provider's business model means that services need not worry about exceeding their idle capacity. More importantly, this means that services don't need to reserve capacity for peak loads. Services should only be concerned about performance during average load periods and reliability during peak load periods. To achieve this, the service should take advantage of other simultaneously running, similar services on the cloud.

There are two different modes that a service can run on the cloud: single-mode and multi-mode. In single mode, a single instance of the service runs on a single machine (for the case of vertical scaling) or multiple machines (for horizontal scaling). While it is possible to run a service on the cloud, in a single-mode configuration, that lumbers along with the load, there are some problems. First of all, due to the high latency networking, it is difficult to build reliable single-mode services. Handling service instances in multiple data centers can be an exercise in futility if the service relies heavily on, for example, an SQL database that has to be replicated to keep data consistent. Most of these services are designed to keep a single-instance service up and running at all times. But this means that peak load service instances must reserve large amounts of resources all the time.

2.5. Resilience in Data Infrastructure

Fundamental services that financial institutions rely on to conduct business often require some level of tolerance for software, hardware, and network interruptions. Even with the proper scalability decisions incorporated into the data architecture, systems will break, whether due to too-high loads, network interruptions, or other only partially predictable failures. Ideally, the effects of such failures would be absorbed by a combination of other, redundant components, allowing the overall system to continue functioning idempotently or allowing the components running in failure to recover and reintegrate without confusing the overall system.

Resilience can be introduced through various common engineering principles, including component redundancy and failover mechanisms. At the simplest level, a component can be duplicated: if one component produces an error or raises a performance flag, the system may automatically inject requests into a second identical component. This design decision is sometimes not well justified. Logically external components, such as currency price feeds from other companies, cannot redundantly be obtained from within the own company. Elements of the infrastructure that are used to store data, such as a central database logistics warehouse, are frequent candidates for redundancy, as they will contain all requested data. However, although replication of this or other storage elements may allow one to tolerate the loss of a single storage component, it is insufficient to ensure resilient service operation. Care must also be taken to interrogate the component consistently, reporting to the user error when the two yield inconsistent results. Repeated data storage at external components can facilitate the resolution of such inconsistencies, allowing repetition of requests toward internal components to attain idempotent operation; however, a permanent decision is required to unblock users and clean data that cannot be repaired. In rare cases, datasets that are often used for the development of business intelligence and analytics capabilities for business teams may contain too much sensitive data: redacting that data may be required to make it available to training and testing environments securely.

Fig 2 . 2 : Resilience in Data Infrastructure

2.5.1. Redundancy and Failover Mechanisms

Financial services companies have traditionally approached the question of how to make their data systems always available, highly performant for query and transaction processing, and also highly consistent with redundancy and failover mechanisms. Such systems today often use SQL databases with available SLAs of more than 99.9999999% for requests per second latency with less than 50ms being a goal. Such systems maintain multiple replicas at low cost usually spread over seven major geographic areas to ensure that they are almost always available even during disasters. Database systems with data and replication growth rates of more than 50% a year and transaction rate growth rates of more than 50% a year rely on replication for read operations to scale. A significant percentage of such companies may have hundreds of billions of queries a day.

Availability is also increased by database systems maintaining a highly available transaction log at high spatial locality for transactions and low serializability wait time during power outages ensuring that recovery from failures can occur quickly. Companies that enhance their query workloads and transaction route distribution with machine learning have increased performance even further. They perform cooling down queries, cache for hot and high locality queries, and shard transactions at a different granularity than queries to the database. Transaction serialization waits, recovery from outages, and infrastructure without keeping redundant transaction log replicas can impact transaction wait times.

2.5.2. Data Backup and Disaster Recovery

An essential component of any resilience plan is creating regular backups of data and maintaining a process to restore system functionality in the event of a catastrophic infrastructure failure. Different systems and data types should be treated differently. These systems and data types may have different backup and retention policies based on the RPO and RTO defined previously. Data types such as user queries to a data warehouse or system logs for a streaming data pipeline may have short retention limits without backups applied. Others may have little to no data backup and long retention policies because storing such data would create prohibitive costs. Whether data is backed up or not, it should be designed with an appropriate loss policy based on RPO/RTO scenarios.

Data backups can be accomplished in very different ways: via snapshots, duplication, change data capture, and others. Nevertheless, the method used should support the data layout and be performant and reliable. Common methods for using backups are recovering full data sets from backups or table merges and appending additional new data from the backup. When planning a backup and DR process, the supporting

infrastructure may also be a consideration in a redundancy path. For example, a malicious actor deleted a system that performed data loading. When the system is restored, may it reload the data set properly? Would it lose data loaded after the last backup was taken? When a pipeline for data loading is manually or automatically repaired after an accident, is it able to reload all data that it has loaded up to the last backup automatically? If not, the pipeline should involve alternatives such as manually loading or rewriting directed files that might take hours to complete. For analytics and warehouse systems, it is essential for business impact when data is missing.

2.6. Technologies Enabling Scalable Data Infrastructure

Most traditional databases are based on the centralized or master-slave model. By contrast, one natural way to build scalable systems is via decentralization – often known as "scale out" systems. New data systems are emerging that are designed to support horizontal scaling fundamentally, without having to resort to partitioning or other techniques that add complexity to the MTDD. Scale-out databases allow transparent replication and partitioning and direct writes to all copies of a record on every page boundary. For Cassandra and HBase, writes to one data copy pass through a commit log, while a background task asynchronously propagates those writes to other copies. Query replication is employed in systems such as Syncauth and Syncoo. Neither of the above approaches protects against loss of active quorum – writes to a quorum without a complete temporary partition can't be applied to returning non-quorums.

Unfortunately, these relatively new models are not the SQL model that is familiar to most IT professionals, and they lack a lot of maturity. Probably the biggest danger in deploying systems like these is their lack of transaction support, especially where multiple records are involved. The oldest replicated databases had no transactions. It is fair to say that without transactions, replicated databases are useless in almost all situations and current systems scale out to such a small number of records as to be not worth considering.

Data warehousing is a technical term of art in widespread use in the computer industry today. It refers to the process of transferring data from many operational systems and consolidating it in the data warehouse, wherein it can be accessed, displayed, and queried using some form of client/server architecture. Typically, the operational systems employ centralized, master-slave type databases – often relational, since the data integrity rules, or constraints, of these systems, are normally quite strong. On the other hand, the data warehouse uses a scale-out database – often a byproduct database, an operational system adapted for the hack data warehouse, or truly decentralized and special-purpose mass storage and retrieval systems.

2.6.1. Distributed Databases

Distributed databases, or databases that can be deployed on multiple nodes, are broadly adopted as the fundamental building blocks of modern internal applications. These systems were originally developed in the 1980s and their main objectives were high availability and fault-tolerance of persistent data. For this section, our focus is on distributed NoSQL databases. NoSQL, which is an abbreviation for "not only SQL," covers all the data solutions that are not in the form of relational databases. In other words, NoSQL databases are schema-less, key-value stores without other relational database characteristics such as normalization, foreign key integrity, and ACID compliance, allowing for a different model for managing and persisting data.

Distributed databases are the working horse for most applications in high-growth companies due to the rapid cycle of how products and services are developed and new features are added. Their schemas are flexible and for most modern NoSQL solutions, the speed of both ingestion and retrieval of data is measured in the microsecond range and horizontally scalable to handle variable workloads. The databases have lower operational costs than SQL-based solutions and most support cloud-native infrastructure with their requirements for containers, management, and orchestration. Since the purpose of distributed databases is to manage persistent application data, more effort has been placed on the consistency, availability, and partition tolerance properties of these databases than on SQL databases.

2.6.2. Data Warehousing Solutions

In addition to transactional use cases, data is required for more analytical purposes to support regulatory and risk reporting, reporting for financial statement close and disclosures, audit, and financial business intelligence. Data from the distributed databases may be updated in batch or in real-time to data warehouses where further transformations happen based on specific rules, running on specific schedules. Data warehouses enable complex queries and reports on historical data. In recent years, due to the significant increase in the volume of data produced, the increase in demand for different data and analytic capabilities, and decreasing costs of storage and computing, organizations are moving towards a multi-layer architecture with a database holding all the raw data, lower costs of storage, and a curated data warehouse hosting larger tables ready for BI teams to run analytics.

In the early days of data warehousing, using massively parallel processing systems, faster and much more scalable queries for discrete timeframes were run against integrated datasets. Then came vendors who helped companies build their enterprise data warehouses, which required heavy upfront investments in on-prem solutions but also

became the single source of truth for enterprise analytics after capturing the ongoing demand for integrated, consistent data from different user groups. Today, existing companies have scaled down on-prem EDW environments, maintaining operational support for business units that have not migrated to the cloud yet. Cloud-based solutions focus on speed, scalability, and simplicity of cloud consumption without the heavy lifting of traditional EDWs. These services allow new players that need to address enterprise analytics faster than investing years in custom-built architecture, to onboard clients and scale rapidly at much lower costs.

2.6.3. Real-Time Data Processing

Data generated in financial services is never static, including transactions, financial positions, compliance reports, alerts, and messages. It changes much more rapidly with events at a lower frequency - a trade is executed, a trade is reconfirmed, a collateral reposting is initiated, or a regulatory report is filed. All of these generate new data value – new reports can be generated, new insights can be gathered or new events that require action get triggered. Financial services also have complex data usage patterns, including ad-hoc requests on historical data and bursts of short-lived activities that use a small set of critical attributes, to long-running batch jobs that use a wide set of data attributes and datasets, such as a security's metadata, position, and transaction history. The newfound ability to gather a range of data attributes and correlate transactional data to other sources of reference data has led to a dramatic increase in new use cases, including surveillance, personalized financial advice, and transaction cost analysis, both within a financial institution and across the financial ecosystem.

The need for real-time processing of financial services data is, therefore, no longer a security facility. It is a basic and integrated requirement, weaving through business processes to create value. The historic batch job mode of processing activity is being replaced by a streaming model – where new insights and new actions are generated by ongoing data processing. Yet, the challenge remains, to scale real-time processing to provide required operations and services that customers expect and regulations require while delivering timeliness, transparency, accuracy, and reliability. Financial services are therefore investing substantially in real-time data services to offer unparalleled levels of customer service and compliance support, and also integrating these into omni-channel platforms to enable seamless experiences.

2.7. Data Governance and Quality Assurance

A robust data governance and quality assurance framework is imperative to ensure the authenticity, consistency, and usability of an organization's data assets throughout its

lifecycle. For financial institutions in particular, the data quality dimensions concomitant with principles of data governance not only need to be adequate for high-frequency decision-making, but they also need to maintain regulatory compliance. At the same time, with the advances in data management technologies, central data governance is also being seen as imposing pipelines of data collection, processing, storage, and use that, at times, lead to organizational bottlenecks. The pursuit of data democratization, which combines self-service access with modular reusable pipelines based on data catalogs has emerged as the potential alternative to the dominant angle of query performance, governance, and standardization.

Before delving into specific guidelines for quality metrics and governance structures for financial institution objectives, it is important to discuss a potential challenge from the growth of a diverse technology ecosystem. Organizations with investments in disparate technologies for the origination, storage, management, and analytics of data are often faced with high costs for implementation, data security, and compliance with regulatory frameworks. Business functions in these organizations that are unable to find appropriate solutions with existing technologies, therefore, often turn to shadow IT for data solutions. Such a situation can lead to security loopholes, high levels of data redundancy, and in extreme cases – costly business failures. It can, therefore, be useful for such organizations to establish a single data ecosystem that efficiently integrates disparate streams of data and metadata across a wide range of requirements and user expertise.

2.7.1. Establishing Data Governance Framework

Data is a key resource in every financial institution. Financial institutions need to properly identify, classify, manage, and protect the data they use for conducting their business. The way data is handled at all levels of the organization is commonly called data governance. While there is no single way to implement data governance at financial institutions, regulators express the need for comprehensive but flexible policies and procedures to be established by senior management and the board. Based on the understanding that data governance is corporately driven, each financial institution's specific data governance program should reflect the institution's own risk profile and needs. A well-defined data governance framework consists of rules and principles for assigning data ownership, accountability, and data stewardship, as well as enterprise-wide standards regarding data protection and usage and department-specific data-related procedures to be endorsed by the data owners. A well-defined data governance framework should also establish procedures and protocols for defining, collecting, storing, aggregating, retaining, and disseminating internal and external data.

Before a financial institution can accurately and reliably report its financial condition, it must know what data it uses; where and how the data is generated, transformed, stored,

and maintained; the system or process controls in place to ensure the validity, accuracy, and integrity of the data; the data subject to any regulatory or contractual restrictions regarding its confidentiality, integrity, or availability; and the contingency plans to ensure availability and reliability of the data in the event of a disaster, cyberattack, or other business disruption. By maintaining a robust enterprise data governance program, the institution will be able to establish a higher level of confidence in the quality of its data used for regulatory reporting.

2.7.2. Data Quality Metrics and Standards

An essential part of a Data Governance program is to define the metrics and standards by which Data Quality will be evaluated. Here, Data Quality Metrics refer to the numerical values ascribed to potential Data Quality Dimensions. For instance, one might compute the uniqueness of social security numbers in a database by dividing the number of unique social security values by the total number of records. Data Quality Standards specifies the cutoffs for the Data Quality Metrics and thus defines which data is considered high or low quality. For instance, the Data Quality Standard might specify that at least 99% of the social security numbers in a particular database must be clean and unique for the data to be of acceptable quality.

While many Data Quality Dimensions and associated Data Quality Metrics are possible, we recommend that banks focus first on the metrics and standards for Dimensions such as Completeness, Uniqueness, Accuracy, Validity, and Consistency, that directly help with detecting common data problems such as missing values, duplicate records, dirty values, outliers, format problems, value problems, and referential integrity issues. These are the Data Quality Dimensions that are associated with the most common short-term Data Quality issues that banks report facing. However, the relevance of the other Dimensions must not be neglected. In particular, the Access, Security, and Reliability Data Quality Dimensions might be of particular relevance to banks, as banks often handle very sensitive Personally Identifiable Information on their customers, such as health records, and they also handle large financial transactions where data reliability is of utmost importance. All of these present Data Quality issues that banks must be aware of and prepare against.

2.8. Integrating AI and Machine Learning

In this chapter, we outline and define how and where to use AI and machine learning when designing a data infrastructure for publishing services. In doing so we distill a practical step-by-step guide. The essence of this work is that you begin with the end in mind, that is to build an infrastructure for data that is amenable to prediction and

therefore most useful for AI-based machine learning pipelines. For this guide, we focus mainly on predictive models, for supervised, weakly supervised, and unsupervised learning.

This chapter is intended for those who are building data learning pipelines for use by AI and machine learning at scale. We outline the details needed to bring these machine learning algorithms into the fold of your ML and AI cloud data infrastructure and might be used by an operation manager, DevOps engineer, or system architect during the implementation phase of a data infrastructure project, working closely with data scientists and machine learning engineers on design choices. The audience may be a product owner of an AI-enabled cloud-serving stack of prediction services or a data engineering manager building the data infrastructure for these prediction services.

AI models can be divided broadly into deterministic models and prediction-based models. Deterministic models such as rule-based code generation, decision trees, and expert systems are systems that are busy coded and do not use data at the prediction stage. They often produce high-quality results and allow for easy inspection of how they work and how the inference is done and are therefore still important tools. However, there have been major advancements in prediction-based models such as AI neural network-based models which have achieved state-of-the-art performance in tasks like language translation, and are even to some extent explainable by indicative rules and heuristics.

2.8.1. AI for Predictive Analytics

Part of the human experience is being able to make predictions about what may happen next, whether that is knowing it will rain tomorrow based on what someone observed about the weather today, or why the stock price of a company went down on a particular day. People make predictions about what other people do or think continuously. A good prediction is based on base rates, which are frequencies observed in the past as to what leads to what, e.g., what kind of people commit murder, and relying on past judge decisions about sentencing people for a particular kind and severity of crime. In business, predictive analytics has been a key component of the management selection process for a long time. In finance, banks and investors have made algorithms based on the past performance of certain companies relative to these companies having a high value-to-earnings ratio or a favorable growth outlook. Further, companies have been penalized with decreased stock price valuations when they exceed or are beneath their known profit levels. AI has significantly advanced the art of making predictions, especially in terms of the period during which an event will happen, thereby improving the management of business risk.

The nature of business is that there are multitudes of quantitative variables that can affect a prediction of corporate performance, e.g., the relationship between price elasticity and demand for a company's product when the company announces a product price advantage is one of known business specialties. In finance, many predictor variables can change to the count of billions when detecting a market anomaly that previously existed when a stock was over or priced at a given period.

2.8.2. Machine Learning Models in Financial Services

Machine learning models are used extensively in the financial services sector. With their ability to predict outcomes based on observed behavior, they are employed for risk scoring, anti-money laundering, fraud prediction, investment recommendation, and propensity to purchase marketing campaigns. Financial services companies have used data to make predictions for a long while; however, the volume, velocity, and variety of new data—transactions, market prices, social media, online reviews, and so on—have changed the kinds of predictions we can now make and the accuracy of predictions made.

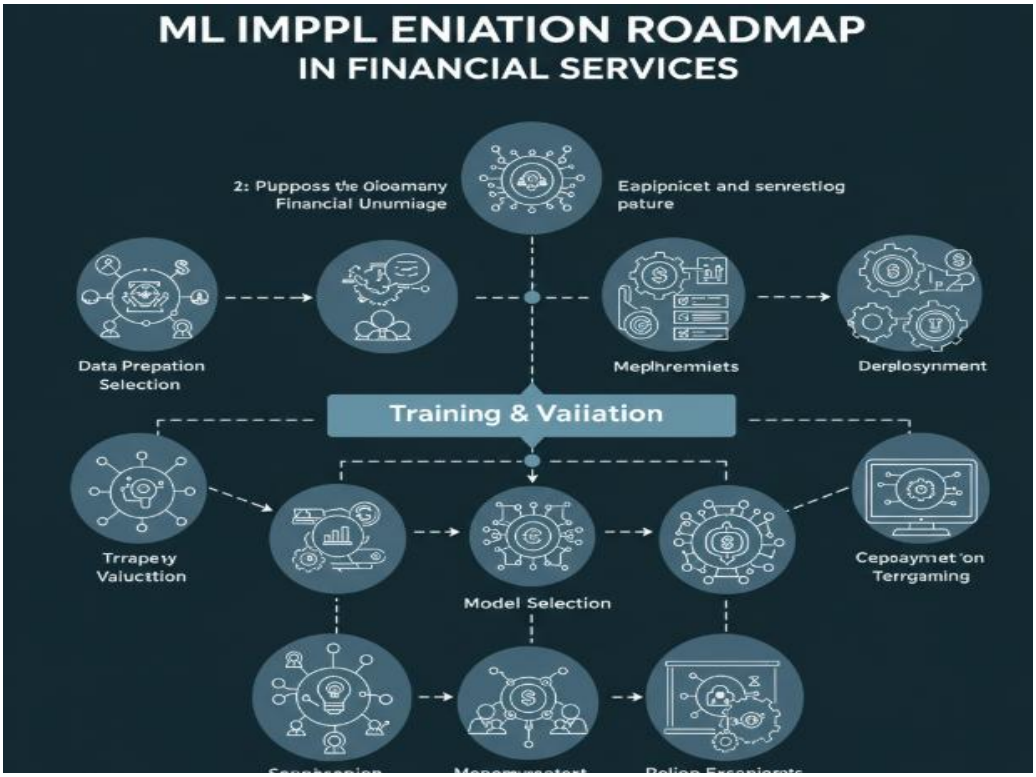

Fig 2 . 3 : ML Implementation Roadmap in Financial Services

Predictive analytics has undertaken exponential growth in multiple facets of financial services. Credit card default models, which fuse consumers' credit history and real-time transaction patterns, have matured to the point where they can correctly predict more than 80% of defaults. Anti-money laundering models, initially based on rule-based systems, are moving to exponential growth via machine learning models, which have lower false-positive rates and improved hit rates. Fraud models have progressed independently as financial service companies have harnessed proprietary point-of-sale networks supplemented by the richness of merchant transaction data. Technology giants, with terabytes of real-time transaction data flowing in from hundreds of millions of users and flexibility in testing, are aggressively pursuing predictive analytics, particularly in the domains of investment recommendations and marketing.

2.9. Conclusion

Final Thoughts and Future Directions in Data Infrastructure

Financial services are now immersed in a radically different landscape than a generation ago. Consumers have unprecedented levels of choice and quality when it comes to services; financial products are being reimagined from the bottom up thanks to raw processing capacity; and delivering those services requires extensive internal resources and complex partnerships, supported by sophisticated tech and data infrastructures. Tech has been de-risked and de-banked; risk has moved into a world of deep and commodity-level cross-border capital flows with enriching tax implications. The old architecture of market-making lies deep in the background of agent-backed decentralized self-custody digital finance; behind the curtains, the scheduling and camera decisions of a wired market hang by a thread over the focus of corporations who now have capital markets playbooks that stretch back to foundational economic principles.

State-of-the-art financial services from the edge of fintech's rent-seeking behavior will be restructured in time-tested but novel ways. As these entities tack deeper into the gulf stream, the available tech supply for core market-facing and capital management services is shifting to the providers who supply both reliability and extreme efficiency. Interest in the layered cloud and modern data stack has only added pressure to these vertically tiered systems—the focus on extreme unit economics; the defense of edge products; and the desire for plug-and-play access to cobbling a core functionality together means that infrastructures in tension across the embedded economy are optimizing for both direction and in-time detail provided to customers. Capability in stability and management of local variables needs to be in place to accomplish the capillarity of approach for systems close-up, which is unique in culture and evolutionary advantage.

2.9.1. Final Thoughts and Future Directions in Data Infrastructure

Data infrastructure is the architecture that enables efficient and resilient data access, storage, and sharing. Due to the burgeoning amount of data flowing into financial services, the infrastructure that serves these needs has internal scaling challenges, is incomplete for increasing complexity and heterogeneity of both source and use of data, lacks the resilience primitive, and is not always compositional, blurring the lines of design intent and approach. We discussed a handful of aspects of the above challenges and the interrelated design tenets of scalability, simplicity, compositionality, and resilience that flow from them, and how they manifest in the decisions we make, from data modeling languages to systems to tools. In addition, we provided examples of the applications that motivate these principles and what happens when they are violated.

Data has become an intrinsic aspect of financial services; the sheer volume and circulation of data have led to the attributions of the moniker of a "data-driven" organization. What started as traditional data access and storage infrastructure has now evolved into a much more complex, hybrid ecosystem with multiple data technologies and tools catering to the needs of multiple use cases. Rigid, traditional concepts of databases and data warehouses no longer capture the inherent nature of data infrastructure today. Cloud computing has facilitated the transition of organizational data services to external vendors, but the goal of delivering timely access to high-quality data for business decisions in a completely scalable and resilient manner remains partially solved. We discuss some design themes and the existing gaps and speculate on possible directions in which the solution space for data infrastructure will evolve.

References

Pfohl, H.-C., & Gomm, M. (2019). Scalable Data Architecture for Real-Time Financial Services. Journal of Financial Transformation, 49, 53–67. https://doi.org/10.2139/ssrn.3480986

Kowalski, M., & Rungta, A. (2020). Building Resilient Data Infrastructures for Financial Applications: Challenges and Solutions. Proceedings of the International Conference on Cloud Computing and Big Data Analysis, 22–31.
https://doi.org/10.1109/ICCCBD48539.2020.9163578

Inoue, T., & Matsubara, N. (2018). Data Infrastructure Optimization in the Financial Sector for Scalable Services. Journal of Computational Finance, 22(1), 51–78. https://doi.org/10.21314/JCF.2018.217

Cilliers, C., & de Waal, R. (2017). Resilient Financial Infrastructures: Key Elements of Scalable Data Systems for Financial Services. International Journal of Information Management, 37(3), 227–234. https://doi.org/10.1016/j.ijinfomgt.2016.12.002

Chan, F., & Mukherjee, S. (2021). Implementing AI-Driven Financial Services: Building a Robust and Scalable Data Infrastructure. International Journal of AI and Data Mining, 7(1), 1–18. https://doi.org/10.1016/j.jaadm.2020.12.006