Deep Science Publishing
https://doi.org/10.70593/978-93-49910-40-9_7

DeepScience
Open Access Books

Chapter 7: Data is the new currency: building and managing financial data pipelines for artificial intelligence readiness

7.1. Introduction

In the past couple of years, various technologies have arisen to automate the building and maintenance of machine learning (ML) pipelines. In particular, the introduction of orchestration platforms and recent research into understanding ML pipeline failures have produced open source tools and guidelines that help users improve the reliability and observability of their ML deployments. These technologies focus on a subcomponent of the end-to-end machine learning workflow, which is the pipeline that produces the ML model, but such a pipeline relies on a similarly complex and critical data pipeline that produces the data used by the ML pipeline. This pipeline typically consists of ETL jobs (i.e., extraction, transformation, loading), data validation, feature engineering, and transformations that must run at various time intervals or data thresholds to keep up with the incoming data for production applications (Moore et al., 2019; Artasanchez & Joshi, 2020; George, 2024).

Managing a data pipeline, however, is significantly more complex than managing a ML pipeline. First, a data pipeline includes many more and more heterogeneous jobs than a ML pipeline. A typical production data pipeline may include hundreds of jobs, compared to a ML pipeline that often consists of three or four jobs. Data jobs can be highly heterogeneous: the job type may vary; the job result may vary; the input data type may vary; and the input/outputs of a job may be both tables and files. Thus, managing a data pipeline requires specialized management techniques that are much more general than those of a ML pipeline .

Second, a data pipeline consists of critical computations that are often hidden from the entities that use the pipeline results. For instance, the ML model lifespan depends on many times larger data and preparation methods than the ML model itself, and its legitimacy and appropriateness rely heavily on being able to verify component exactness and performance (not just infeasibility). Thus, breaches of public trust in data pipelines lead to larger and sometimes insurmountable issues that an organization must face to continue using the flawed products (Polak et al., 2020; Stadnicka et al., 2022; Syed, 2025).

.

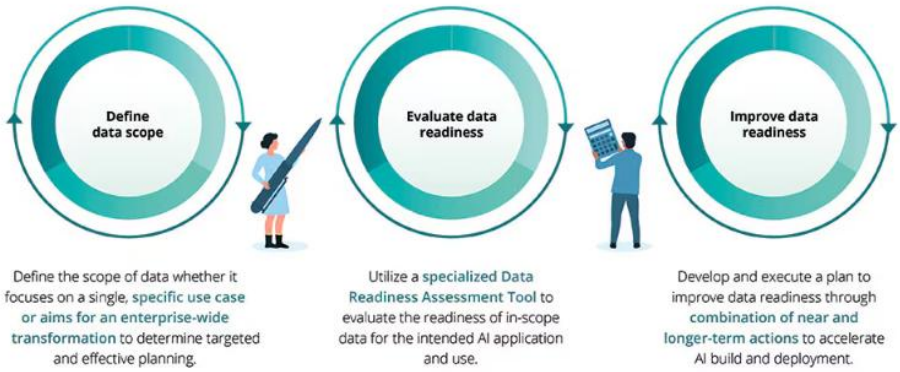

Fig 7.1: Transforming AI Outcomes with Effective Data Readiness

7.1.1. Background and Significance

Data has always been a critical asset for organizations of all sizes and forms. Data lacks any intrinsic value, however. The value that people place on data and the business models of firms that deploy data-driven applications largely depend on the varied ways data is turned into useful information. Corporations and organizations actively build new data resources and collect data as a byproduct of running their day-to-day business processes. Such data, created or owned by an entity, is typically termed private or internal data. Organizations enter the data marketplace for two main reasons. First, they may need to complement their existing data resources or fill the data gaps needed to train a successful machine learning application and cannot wait for the ever-expensive and tedious data collection process. Organizations may wish to monetize their private assets by packaging or enriching them for external use, such as placing sensor data into the marketplace for third parties to view analytic trends rather than simply supplying the raw data itself.

Translating data into wealth generally cannot be done simply by creating or buying data. The processes of cleaning, curating, preprocessing, and transforming the raw data into a more suitable format, as well as tightening various aspects of data governance and reliability, come into play. Organizations aiming to graduate from raw data to mature data products or tasteful data products need to engage in data engineering. At the core of data engineering lies the data pipeline, a series of computational processes through which the raw data product is transformed into a refined one. A data pipeline can take various forms in terms of topology and scope. In simpler terms, it is not a pipeline but a complex system consisting of a collection of interconnected tasks, which ingest data from diverse sources, process data in various fashions using varied computational paradigms, and output data resources of different types and with different qualities.

7.2. The Importance of Data in Finance

The importance of data in finance cannot be overstated. Financial data, often characterized as currency in world finance, plays a significant role in determining an organization's fortunes. Growing volumes of data continue to pour into the web and enterprise systems, including unstructured data hidden away in files. New streams of unstructured data continue bubbling up from smart devices, the Internet of Things (IoT), streaming social media channels, and Web 2.0 applications. Smart web crawlers let enterprise systems grab whatever information is needed to develop a competitive edge. Natural Language Processing (NLP)-enabled technologies enable organizations to unlock insights out of unstructured data present in audio, video, documents, and press releases, hence enabling better decision-making.

Intelligent systems put down roots in finance. Data and its analysis in tandem with ever-increasing computing power lead the financial industry into a golden age of data-driven solutions—driving profits in algorithmic trading, asset risk prediction, fraud detection, and better 'differentiated' pricing. Over $865 billion is poured in by Tech Giants and increased revenue figures in Data Science/Big Data businesses show that a tidal wave of data is owned by them. And yet, there is still a tremendous need for capital funding to get insights out of financial data generated by unregulated social media platforms, dark web chatter, IoT-enabled customer sentiments, and web scrapped company news, hence preventing crises like the collapse of Silicon Valley Bank.

7.2.1. Research design

The research focus is to analyze the technical requirements for operating a financial data pipeline and converting the financial data ontology into trustworthy, clean, and conveniently accessible data. All the data on exchanges, wallets, and block explorers is

available on the web for free but difficult to access. In the financial industry, the terms datasets and data pipelines are often confused but should be treated separately. Financial datasets refer to a finite collection of data points that are stored permanently in a vault or warehouse. Financial data pipelines refer to the flow of live data from one or multiple sources to the end consumer and are typically in or close to real-time. This paper offers detailed insights into building a data pipeline for cryptocurrency-related fundamental data. An example of such a data pipeline for mining-related wallets is demonstrated. This data pipeline aims to adhere to the open data philosophy, where all the data can be freely accessed and reused without restrictions while complying with industry standards.

An analysis of the asset/market structure in the cryptocurrency industry and an accompanying data model are derived. Additional insights are provided into the technical hurdles of implementing such a data pipeline and the respective solutions. One of the major hurdles of the current deliverable is the on-chain data's sheer amount, structure, and privacy. Technical requirements regarding data indexing, financial ontology and parsing, data transformation and cleaning, and user interface and accessibility are proposed and elaborated . Data pipelines, in this paper, refer to the ETL processes taking place within a financial entity to guarantee an excellent quality of standard financial datasets, which will move on to vaults or warehouses.

7.3. Understanding Financial Data Pipelines

Data has always been a critical driver of business success, but the financial services industry is more heavily dependent on good data than ever before. From regulatory compliance to smart investment processes, AI & Big Data initiatives would be impossible without proper financial market data. This also brings challenges, as facing new vendors and ingest sources is now a daily obligation. Data preparation can take weeks; therefore, a data pipeline is critical for data takeovers and scaling up for use cases. Data is the new currency, and managing financial data is now just as critical as managing large sums in accounts to lose money.

Financial pipelines are the set of steps required to prepare data, enabling a successful exchange between data sources and data consumers. To automate the interaction with data sources, a sequence of stages is applied: Ingest → Source → Scrub → Transform → Load → Distribute. Speed requirements dictate whether to consume snapshots for up to two-week-old data or to stream inter-day updates on the fly. To cater to needs, the ingestion methodology must be adapted to the size of the datasets and their update frequency.

With emphasis on the set of stages required for data scrubbing, fundamentals of time-series databases (historization) and the benefits of columnar databases are highlighted.

Data is the new currency; therefore, efficiency and consistency-oriented data consistency algorithms guarantee that accurate data is always seen by consumers, even in the face of failed exchanges and network partitions. Financial data is mostly time series; therefore, a scalable windowed approach to smooth nourishment without vast amounts of redundant data is proposed.

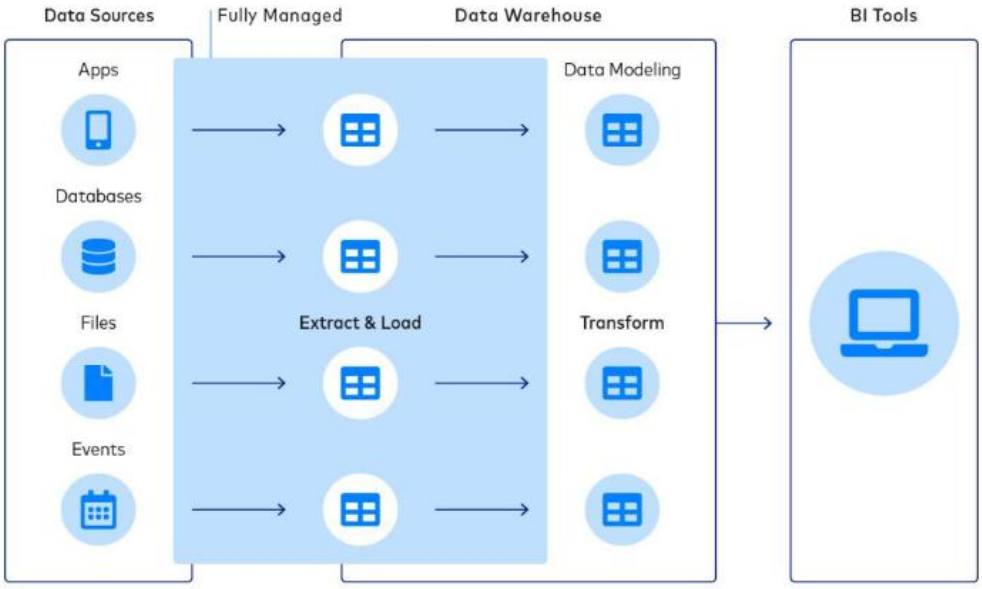

Fig 7.2: Financial Data Pipelines

7.3.1. Definition and Components

As businesses and governments strive to turn data into a new type of currency and draw the right conclusions from it, the need for high-functioning, financially-managed data pipelines has emerged. Existing big-data pipelines are incapable of supporting mature AI-based business models. They cannot find "nuggets of gold", and they lack standard cost management and financial metrics. Building on the concepts of financially-managed data pipelines and data-liquidity metrics for big data platforms, this proposal for a new data architecture for AI pipelines leverages active machine learning based on new conceptual models of data pipelines along with their data liquidity metrics, accounting and finance practices, and visibility-enhancing data curation processes. Finally, this architecture and metrics offer nonprofit and regulatory governance approaches to make data all stakeholders' business infrastructure and preserve democratization of AI.

Financial Data Pipelines refocus attention on bigger picture performance targets than those associated with individual data quality and operational characteristics, and they enable enhanced executive visibility. They shift attention from raw-databases-on-bit-

buckets to manage-on-data themes/units, rather than traditional design component-based operability. They retain good properties of current data pipelines but also leverage more sophisticated, target-formatted data curation and tasks. Traffic to theme pipes is elastic and throughput time adaptable to data production variations, and queues are efficiently managed. Good theme data quality emerges from appropriate time-aligned systematic data curation, ensuring best accessibility/consumability. Finances govern theme data ownership by either business or regulatory stakeholders.

7.3.2. Types of Financial Data Pipelines

The goal of Financial Data Pipelines (FDK) is to connect existing financial data sources to stakeholders, including firms, regulators, and clients, for immediate use in Various Non-Traditional Models (NTM) initiatives, such as financial data re-use in digital transformation, fintechs to support open banking initiatives based on regulatory fulfilment, and enhanced connectivity between distributed financial services systems to exploit the potential of highly granular financial market data. To alleviate duplication and efforts in the build processes of existing financial data sources in Information Technology (IT) systems at firms/regulators, FDKs provide storages, transfer/protocols engines and Artificial Intelligence (AI) tools for easily connecting data policies and archiving state of data post-execution, to control costs and risks involved. FDKs comprise four components: the physical, the forensic/accounting, the structural, and the contextual. The physical component is concerned with designing the storages and methods/protocols for transferring financial data as financial instruments and executing NTMs. This component is the most complex and incorporates data technologies, and AI/ML tools selected via the technological fields/taxonomy proposed. The forensic/accounting component design involves a catalogue of key metadata of financial data sources, including time-stamps of last updates, the executed NTMs, and the stored archived states of financial data for each executed NTM. The structural component is concerned with searching similarly structurally designed storages. The contextual component identifies reasons for the existence or absence of financial data at each financial data source with a current state.

FDKs for existing regulatory and firm financial data sources entail a careful build process of these components. FDKs for firms/regulators entail 1. Identifying and feeding the forensic/accounting component, and 2. Building the physical component, in all cases.

7.4. Building Financial Data Pipelines

This section presents the life cycle of building and managing Financial Data Pipelines. It starts from identifying the business metrics of interest, involves digging down to the

data metric aspects, and finally building the Financial Data Pipelines (FD pipelines). FD pipelines are implemented by setting up the source-to-target data flows either based on a rudimentary ETL tool, cloud industrial-grade data pipeline tools, or orchestrated by workflow engines. Then it covers managing the FD pipelines, such as monitoring data quality at the source, and creating slide dashboards reporting the FD pipelines status. In the end, mitigations of data spikes from upstream systems and long-range cascading data downtimes are discussed. Establish a seamless Financial Data Pipeline between upstream systems and targeted Business Intelligence (BI) reports for Cartesian data metrics and driven by the required business knowledge. Given the vastness of data sources and target systems raise the challenge for gathering information initially, clustering data aspects related to each other based on speakers' descriptions helps sift through potential metrics. Allocating separate offshore development teams for well lagged BI solutions frees time for more pressing data flows or new slowing data laps classes explosively growing data. Backstopping the offshore teams with stage gating of deliverables ensures reproducibility. Using a combination of ETL/ELT/BDP tools provided by big data cloud vendors reduces the local maintenance cost. Given the ballooning number of source tables restructuring quarterly weight can provide up to three months of headroom. Drawing out simple source data transformation documentation helps greenset up their bespoke environment.

7.4.1. Data Sources Identification

The first process of preparing data for large-scale AI involves the identification and exploration of data sources from within and outside the enterprise. This should include a thorough catalogue and description of all data sources potentially of interest. If necessary, data quality can also be established. With vast amounts of streaming data from a plethora of sensors, devices, logs, and business transactions generated and stored every day, the popularity of big data technologies appears to be well justified. Big data lakes seem to be the answer for these potential data-mining gold mines. Research continues on how to best design a big data lake and how to pair state-of-the-art and yet to be perfected analytic tools with cost-effective data storage technologies. However, whilst large amounts of data remain stored, many companies face a situation where no or only suboptimal data are used for valuable analytics. To date, only traditional data warehousing and OLAP tools seem to have been applied successfully to mission-critical transactions. Enterprise data lakes have not yet proven sufficient for actual and continuous high-quality insight extraction rates. Besides this unsatisfactory latency, traditional IT architectures seem to struggle with gaining new sources of data, such as semi-structured and unstructured text produced in either, or both, on- and offline settings that were previously impossible to keep. This is an issue since this data seems most promising to deliver added-value insights. Furthermore, it is expected that up to 80% of

future sources of big data stocks will be non-structured data from a plethora of technologies, devices and channels. This suggests an inadequate architectural structure to current and future business needs. It seems time to redefine the enterprise's data architecture. Summary of earlier research is thus required as a prerequisite to a new proposal for an enterprise data architecture. Research within the scholarly community is classified by three categories useful for the review of outcomes. The first is ontologies, standards and so forth aimed at characterising different types of structured data within an overall architecture close to Moore's Architecture of Destruction. Unfortunately, these contributions seem outdated and incapable of solving currently encountered business needs. The second field is enterprise data management followed by many variants, mostly aiming at service-oriented, data-centric and client-centric design principles. The need is visible in data-driven markets with clients expected to connect anytime, anywhere. In these markets, internal functionality-oriented silo interactions are no longer competitive; data that travels with clients should suffice.

7.4.2. Data Ingestion Techniques

Data ingestion techniques are elaborated in this section. Ingestion is an important activity that is performed when loading new raw data into data storage systems and is typically the sole responsibility of the data management system. However, complex computations are typically applied to surface data in consumption systems for analytics and other OLAP workloads. These computations can also take the form of ingest-time enrichments, where a part of the operations is pushed from the consumption system down to the ingestion pipeline fronting the data storage system. To support these ingestion enrichments, the ingestion pipeline needs to be adapted to run the specified operations as part of the ingestion process. To adapt to link changes, it needs to be able to add, remove, replace, and modify or change operations based on various events. An ingestion pipeline is presented, which allows end-users to define ingestion pipelines and pairs of storage and ingestion systems. The ingestion system could be instructed to run the ingestion pipeline by the storage system. The ingestion system polls the defined ingestion pipeline of the storage system and looks for new ingestion jobs displayed on it. A new ingestion layer (IDEA) is proposed on top of AsterixDB, which implements schemas for an ingestion pipeline and a set of source control statements.

AsterixDB’s native ingestion system ADAM plays an essential role in blocking-based and streaming ingestion. In addition to the current capabilities of ADAM, a compile-time strategy to enable ingestion enrichments requiring non-trivial operations and an automatic rewrite strategy for handling changes in the input sources can be realized. Whether freshness or correctness is guaranteed, in addition to the current waiting behavior of ADAM, a timestamp and event record are needed to be tracked for the

ingestion job pages of external inputs and reference data changes. AsterixDB's data enrichment APIs can be modified to adaptively maintain newly bound acquire materializations. The design expands AsterixDB's current ingestion workload by allowing operations of enriched rich raw data to be performed, thus enabling more complex applications with complex enrichments.

7.5. Managing Financial Data Pipelines

Managing financial data pipelines has its own challenges. To improve data pipelines for AI readiness, designing, extending to the cloud, and observing day-to-day management need to be considered. Pipelines must adhere to a standard structure, where configuration should be designed distinctly from implementation. When constructing or modifying a data pipeline for an existing use case, a wizard should guide the design of the pipeline structure. Provisions should be made for cloud infrastructure as well. To observe data criteria over a cloud-native setup, building infrastructure would not be reasonable; instead, usage must be planned as a service (ideally with open connectivity). This means it is necessary to adhere to a flexible format standard for configuration as well. On the PCF, standard configurations serve as a potential check before using the cloud storage pipeline. In general, managed services and prepackaged connectors are highly desirable for the flexibility to extend the cloud data pipeline. It is unnecessary to throw or convert the data back and forth when using similar public cloud storage or databases. Whether in the cloud-Native AI pipeline, data should be uploaded from the edge directly before consideration. If the data comes from a pre-filtering center or there are potential data loss issues, the first dataset to use should be planned thoughtfully, as the steps for data ingestion won't be easy to change later. Thus, log collection should initially use smaller tokens while ingestion copies larger text tokens. To analyze potential problems, design should iterate on observability, which is rather an explicit case. Data timeliness and currency could be checked one after the other, respectively from periodic checker and duplicate data size. Skewness is better to analyze storing with p.blue but obfuscated textual tokens stored with p.buy and p.sell are advisable. Ad-hoc analytic needs to extend cloud-native but analyzing outside of the data lake should be pre-planned as likelihood of its potential increase is rather harder to shut down.

There are several trends regarding the future of data. One is that data will follow the same trajectory that many other sectors have: large-scale transaction networks of producers and consumers will emerge, creating vertical integration in the data market and developing liquidity, transparency, and timeliness of pricing and data quality across the market. Much like the lag between similar networks emerging in stock trading, banking, and other sectors, it will take time for these features to become ubiquitous across the data market, but data is heading in this direction, especially in sectors deemed

'data poor' for neither technical nor regulatory reasons. In these sectors, there is a clear advantage to integrating the data value chain and employing advanced trading mechanisms.

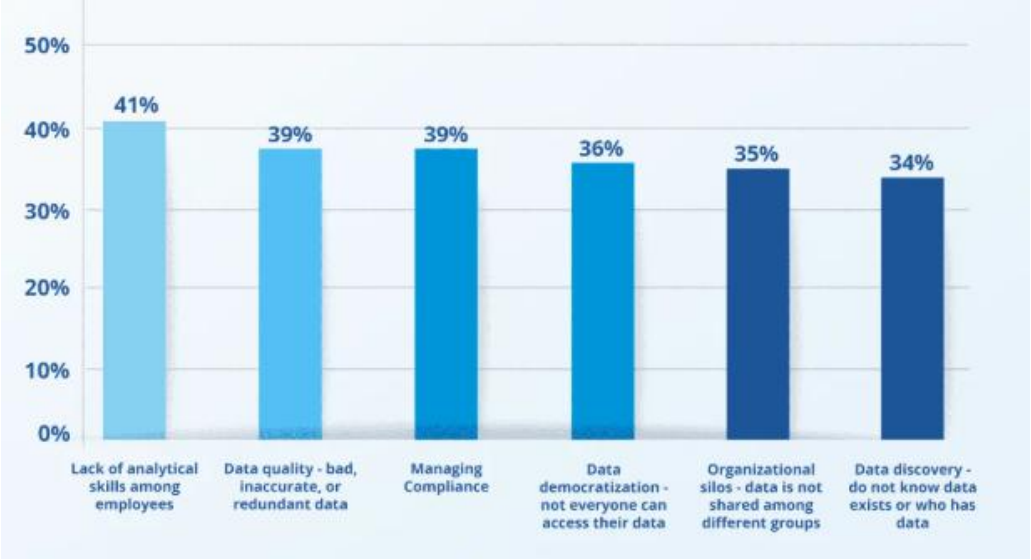

Fig : Data Preparation

7.5.1. Monitoring and Maintenance

Despite the extensive work on model monitoring and maintenance in general ML use cases, this still remains an open challenge for ML models deployed within firms. Research has highlighted an undocumented gap between the development of a data-driven algorithm in the lab and its deployment within a firm. On the one hand, this affects the usability of current models in production that were trained or calibrated on purely historical data. On the other hand, it also creates opportunities for research that can tighten the loop between academic model development and the ML practice of firms. Baby steps would be to work closely with practitioners with domain expertise in a specific area of analysis and who can help develop the appropriate models or algorithms that suit the firm's needs better.

A relevant topic to ensure AI-readiness and therefore wider adoption of existing models is addressing data drift. Current DR techniques within firms typically focus on prediction drift, as this is often readily available and interpretable to offer model insights to non-experts.

7.5.2. Data Quality Assurance

The successful development and implementation of a data pipeline is still no guarantee of project success. Once integrated with business processes, the pipeline is bound to supply new data, which means that all pipeline components are under continuous strain to ensure that this data is both useful and correct. It should be expected that new issues subsequently arise, such as one-off pipeline failures and one-off problems with the data, but also continuously recurring problems with correctness or usefulness. Automated tools for quality checks and correctness assessments should be an integral part of any data pipeline for AI systems, just as tests and unit tests are for production code in general software development.

When it comes to machine learning or AI, data quality is a broader term. It includes the aspects of correctness and usefulness, of course, but also includes facets such as bias and even privacy. Except where otherwise noted, the focus is on correctness and usefulness here. All the verbal specifications must be fleshed out technically, that is, all the things that can be checked about the quality of data must be formulated in such a way that they can be automatically checked by a machine with no human intervention. For this, the templates constructed earlier can be used, enhanced with details about exact threshold values, appropriate machine learning models, or possible corrective actions. A selection of approximately one hundred quality checks can be found.

7.6. Conclusion

Building financial datasets and pipelines is a multi-faceted task, which demands managerial, technical, and organizational know-how from different areas of expertise. In this chapter it is argued that the topic of financial data, comprising sources, formats, structures, management, pipelines, and market implications, is in many ways similar to the classical topic of currency by presenting its own “currency” framework comprising a number of important questions. In recent years, the topic of currency has been addressed through theoretical essays steeped in economics and policy considerations, and through practical works focusing on the view of different stakeholders engaged with cryptocurrencies, but a systematic treatment of data as a currency capable of providing an equally broad and diverse view is lacking.

The topic is significant for at least three important reasons. First, more financial data than ever is generated, collected, and produced by more devices than previously thought possible. Second, and as a consequence, the costs and overheads involved in building data pipelines have surged. And third, the increasing power of AI and machine learning models natively relying on complex data accesses, and the growing market infrastructure hitherto only offered for numerical currencies. It is anticipated that the topic of currency

will continue to provide fertile ground for innovative thinking and debate and that the subject of data in finance will offer many new areas of knowledge and opportunities.

Building datasets and pipelines from the existing sources of financial data is not the gift of a single person but a collective effort of professionals from different domains. The data engineering or data preparation stage of an AI or machine learning project is an extensive undertaking, which typically involves the collaborative efforts of data engineers, making decisions on data sources and formats, data quality issues, metadata schemas, cleaning and transformation processes, summarization, and exploration procedures. But the same can be said about the governance of a cryptocurrency, with key managerial practices and organizational decisions on which networks, blocks, nodes, wallets, smart contracts, authorities, and chain forks to adopt. It is important to capture and systematize this essential knowledge which allows for the construction of financial datasets and pipelines.

7.6.1. Emerging Trends

As there is rich temporal ordering in many data types—especially commercial and social data—data exchanges may also develop new tools for time series databases, providing liquidity across time-derivative instruments such as quarterly earnings, v-shape predictions, and more. Some first movers in the data market have begun to develop sophisticated trader environments for temporal data types, but a wide disparity exists between users with early access to these capabilities and the majority of users that do not have them. Additionally, with vast amounts of data flowing in from IoT devices, a new class of networking/communication architecture is being developed. Edge computing is gaining traction, wherein transactions are executed on resource-constrained devices that may also be mobile, increasing the demand for various ways of managing quality, fidelity, and accuracy in the modern economy based on use-and-return data models. Some of the above trends will happen faster than others, and while both cultured markets and edge architecture can have a significant impact on providing data liquidity, they will take longer than the less complex trajectory of manufacturer-device exchange.

References

Polak, P., Nelischer, C., Guo, H., & Robertson, D. C. (2020). “Intelligent” finance and treasury management: what we can expect. Ai & Society, 35(3), 715-726.

George, A. S. (2024). Finance 4.0: The Transformation of Financial Services in the Digital Age. Partners Universal Innovative Research Publication, 2(3), 104-125.

Artasanchez, A., & Joshi, P. (2020). Artificial Intelligence with Python: Your complete guide to building intelligent apps using Python 3. x. Packt Publishing Ltd.

Moore, J. H., Boland, M. R., Camara, P. G., Chervitz, H., Gonzalez, G., Himes, B. E., ... & Holmes, J. H. (2019). Preparing next-generation scientists for biomedical big data: artificial intelligence approaches. Personalized medicine, 16(3), 247-257.

Syed, S. (2025). Machine Learning Algorithms for Optimizing Big Data-Enhanced Cybersecurity in ERP Ecosystems. Journal of Artificial Intelligence and Big Data Disciplines, 2(1), 36-44.

Stadnicka, D., Sęp, J., Amadio, R., Mazzei, D., Tyrovolas, M., Stylios, C., ... & Navarro, J. (2022). Industrial needs in the fields of artificial intelligence, Internet of Things and edge computing. Sensors, 22(12), 4501.

Socol, A., & Iuga, I. C. (2024). Addressing brain drain and strengthening governance for advancing government readiness in artificial intelligence (AI). Kybernetes, 53(13), 47-71.