Deep Science Publishing, 2025
https://doi.org/10.70593/978-93-7185-061-2_4

DeepScience
Open Access Books

Chapter 4: Architecting Cloud-Native AI Ecosystems

4.1. Introduction to Cloud-Native AI

Cloud-native AI describes an emerging approach to creating and deploying artificial intelligence capabilities alongside the emerging approaches and best practices of cloud-native software development, which itself is following and expanding upon the DevOps and Agile lines of thought. Cloud-native AI combines the power of scalable storage and compute capabilities with cloud-native software principles and best practices in order to create AI models and to deliver AI functions to end users.

Conventional (or “non-cloud-native”) AI development platforms and ecosystems are characterized and driven by a number of factors: relatively slow and therefore relatively costly model-building processes; monolithic services that combine many functions and that are deployed within the same process; the use of virtual machines to deploy services; complex and monolithic production environment constructs; the use of platform as a service (PaaS) renditions such as IBM Watson; and the general absence of AI development tools—services and pipelines—into the AI model-building process. Cloud-native AI employs cloud-native design principles to minimize and often to eliminate these external constraints. Taking advantage of containerization, microservices hosting platforms and orchestration frameworks such as Kubernetes, serverless compute, and DevOps-based AI model and AI service development guidelines, Cloud-native AI—typically deployed in AI developer laboratories—enables significantly faster and less costly AI model-building activities. The APIs and containers of the cloud-native AI ecosystem can then be made available for deployment in the more productionized environment of the AI model execution environment.

4.1.1. Overview of Cloud-Native AI Concepts

Cloud-native architecture deliver AI application platforms capable of use-oriented system construction. Cloud computing models accordingly designed for the seamless

incorporation of AI technologies support automatic high-level deployment and orchestration of AI-specific cloud resources and AI application components, such as pre-/post-processing and training services. Data centers form the foundational infrastructure hosting cloud resources, leveraging cloud platform services to facilitate the definition, creation, and updating of cloud-hosted AI applications that meet typical use case requirements.

The integration of AI technologies into cloud computing frameworks—cloud-native AI—enables the direct provision of superior AI-capable cloud services. Cloud-native AI follows an AI-centric design pattern that exploits the inherent characteristics of cloud platforms when building AI application services. This approach encompasses the development of AI-native cloud infrastructures endowed with neural accelerators and protocol architectures tailored for AI operation. Additionally, cloud-native AI supports various cloud services incorporating machine learning, including automatic model training, model serving, pipeline scheduling, and AI model marketplaces.

4.2. Fundamentals of Cloud Computing

Cloud computing is the delivery of computing services such as servers, storage, databases, networking, software, analytics, and intelligence over the internet ("the cloud") to offer faster innovation, flexible resources, and economies of scale. Cloud services are typically classified into three types depending on the level of abstraction or, essentially, the management scope: infrastructure, platform, and software. Infrastructure as a service (IaaS) provides virtualized hardware instances, which can be used to host software in VMs. Platforms as a service (PaaS) enable the provision of a deployment environment on top of the hardware instances, whereas Software as a service (SaaS) provides access to an actionable or functional deployment for end users.

A cloud computing setup can be built as a public, private, or hybrid cloud, depending on its deployment scope and targeted clientele[1-2]. A public cloud provides shared or dedicated cloud-management services to the general public, a private cloud provides shared or dedicated services to a single organization, and a hybrid cloud integrates both environments.

4.2.1. Types of Cloud Services

Cloud computing offers a variety of services that deliver computation, software, data access, and storage as utilities, accessible when needed. These services promote users' independence from specific infrastructure components through a flexible pay-as-you-go model. Cloud services combine and virtualise aspects of traditional data centers—such

as networks, servers, storage, applications, and services—by employing virtualization, distributed computing, and resource management technologies. Four core types of cloud services exist, alongside two cloud deployment models: public and private.

Cloud Software as a Service (SaaS) provides users with access to infrastructure-independent applications like word processors and online backup programs. Cloud-based Application Platform as a Service (PaaS) offers support for applications, including operating system services and databases. Cloud-based Platform as a Service (PaaS) encompasses control, operational, and middleware services that support the development, deployment, and operation of SaaS applications and mobile applications. Cloud-based Infrastructure as a Service (IaaS) delivers managed and Internet-based access to servers, storage, and networks, enabling the operation of a platform environment.

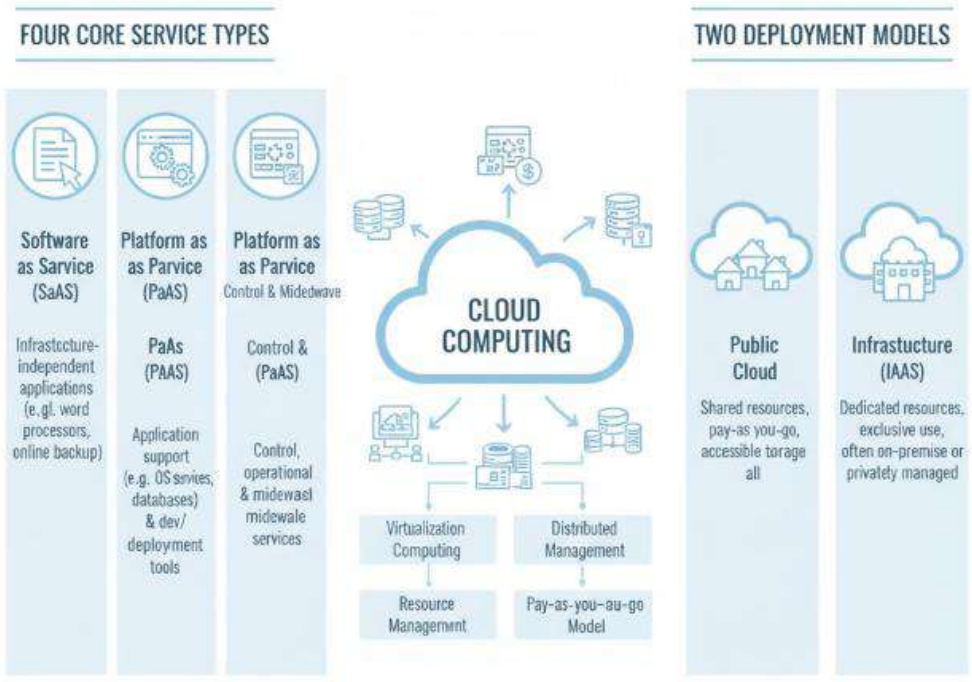

Fig 4 . 1 : Cloud Computing Services and Deployment Models

4.2.2. Cloud Deployment Models

The main cloud deployment models—public, private, community, and hybrid—define where and how services are implemented.

Public clouds are available for use by the general public and are most often owned by a cloud service provider. Access to public cloud services is normally allocated on a pay-

per-use basis. Private clouds are operated solely for a specific organization. They may be managed by the organization or a third party and may exist on or off-site. When private cloud computing infrastructure is shared by several organizations and supports a shared mission, it qualifies as a community cloud. Hybrid clouds are specialized clouds that utilize two or more cloud deployment models in combination, such as a private cloud linked to a public cloud, enabling data and application portability.

4.3. AI Technologies and Frameworks

Foundations of artificial intelligence techniques and frameworks enhance the capabilities of cloud infrastructure and services. Machine learning, natural language processing, and deep learning constitute key approaches that enable intelligent applications and services.

Machine learning algorithms utilize statistical methods to enable computer systems to progressively improve at specific tasks with experience. These methods encompass supervised, unsupervised, semisupervised, reinforcement, self-supervised, and ensemble learning, and apply to areas such as computer vision, natural language processing, and autonomous systems development. The scikit-learn and Spark ML libraries offer commonly used implementations. Deep learning, while a subset of machine learning, employs artificial neural networks—typically with more than two hidden layers—to perform tasks that often require the establishment of complex abstractions. Implementations include Theano, PyTorch, Keras, and TensorFlow, with specialized architectures such as CNNs, RNNs, transformers, and GANs. For natural language processing, essential building blocks such as tokenization, lemmatization, part-of-speech tagging, named entity recognition, and dependency parsing are provided by frameworks including NLTK and SpaCy.

4.3.1. Machine Learning Basics

Machine learning enables computers to identify complex patterns in data and learn directly from examples—similar to how humans learn. This ability allows computers to tackle problems too complicated to solve with explicit instructions. Over the past decade, deep learning methods have dramatically expanded the capabilities of machine learning systems. Unlike traditional, rule-driven computer programs, machine learning systems can gain new skills autonomously by adapting to new data and scenarios.

As a result, modern Artificial Intelligence applications often produce smart behavior, dialog, or speech that approaches human quality and thus seems ‘‘intelligent.’’ However, since these systems do not detect or learn new concepts using the same processes as the

human brain, the question ‘‘Can machines think?’’ jumps from philosophy to science fiction [1-3].

4.3.2. Deep Learning Frameworks

The field of artificial intelligence witnessed a paradigm shift from explicitly programmed expert systems to neural network-based techniques, highlighting the prominence of machine learning as a core AI approach. While shying away from classical pattern recognition methods, recent neural network applications address recognitions tasks in vision or speech. Tools like Open CV provide rich libraries of computer vision algorithms useful for pre- and post-processing phases, with integration into cloud-native environments facilitated through the cloud provider’s APIs. Flashlight and Librosa offer support for fundamental audio signal processing operations such as audio feature extraction, enabling feature engineering for speech applications.

Deep learning excels in both unimodal and multimodal pattern recognition and generation processes, which are central to many AI applications. Clouds have become the predominant AI execution environment for emerging pattern recognition applications such as textual, acoustic, and visual recognition and generation. Classic models of unidirectional recurrent neural networks such as long short-term memory networks or gated recurrent units evolve in state-of-the-art transformer-based models. Public cloud resources, including specialized hardware accelerators like Google's Tensor Processing Unit, allow for scalable training and deployment of these extensive models.

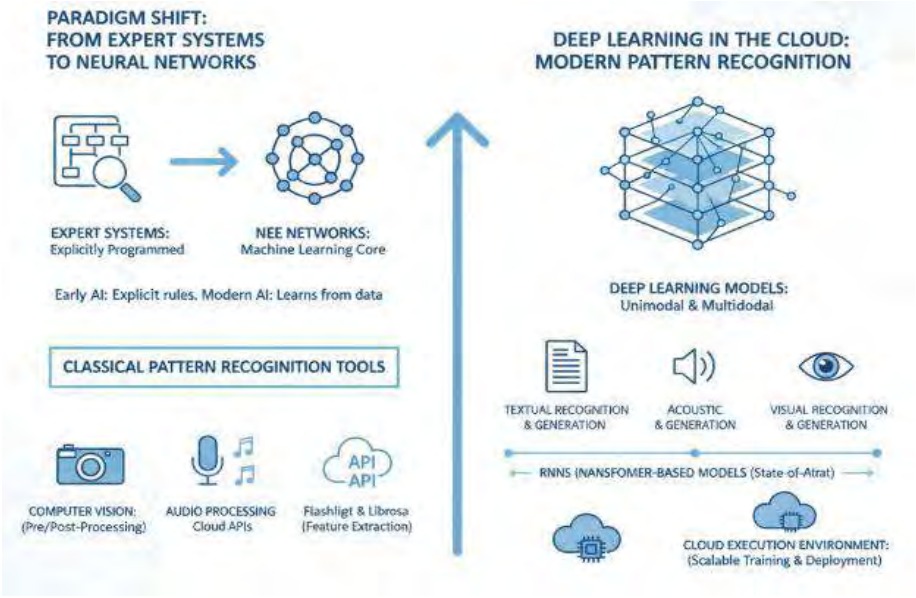

Fig 4 . 2 : Deep Learning, Pattern Recognition, and Cloud Execution

4.3.3. Natural Language Processing Tools

Recent years have seen a plethora of natural language processing (NLP) toolkits and text-processing services. They range from generic toolkits designed for multiple languages to more focused options that cater to particular tasks, such as coreference resolution or sentiment analysis. Other toolkits offer APIs that simplify access to URL extraction and text classification features.

NLP service providers, on the other hand, are big players who have constructed APIs that offer a plethora of text-processing methods. Microsoft, with Text Analytics, Google, with Natural Language API, Amazon, with Comprehend, IBM, with Watson Natural Language Understanding, and Aylien, with Text Analysis API, are all competing for market share. Deep learning frameworks, such as BERT, not only provide pretrained NLP pipelines for a variety of tasks but also serve as pretrained models for transfer learning, which can be used to create pipelines with high-quality results in low-resource scenarios. Keyphrase extraction is one of these tasks: Document and extract keyphrases from short or lengthy documents.

4.4. Design Principles for Cloud-Native Architectures

The transformational impact of Artificial Intelligence (AI) on digital businesses is profound, as reflected in eBay's commercial adoption of AI for efficient bidding and Yahoo's utilization of natural language processing to automatically classify news headlines [2,4-5]. The cloud computing paradigm, with its rich ecosystem of infrastructure, platform, and software services, provides unprecedented agility and dynamism for developing and deploying AI applications. The convergence of AI and cloud computing is indeed invaluable. Cloud-native architecture, a new version of cloud architecture, embodies concepts such as declarative programming, parallelism, modularity, CQRS, ES, scalable systems, and microservices. These architectural principles underpin the exploration of ecosystem architecture methodologies for cloud-native AI, encompassing microservice design, container orchestration, data storage, pipeline design, regulation, and security.

Cloud computing delivers hardware, software, and other resources as services over the internet, requiring minimal client maintenance and offering superior adaptability and resource management. It enables organizations to access high-powered computing infrastructures, platforms, and applications on a pay-per-use basis. The essential characteristics of a cloud include on-demand self-service, broad network access, resource pooling, rapid elasticity, and measured service. Deployment models comprise public, private, community, and hybrid clouds, while service types are categorized into Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a

Service (SaaS). Cloud-native design entails building and operating applications exploiting the cloud model, with an emphasis on speed and flexibility. Cloud-native techniques facilitate acceleration and reliability in the development, deployment, and operational phases, enabling organizations to generate business value faster. Although not inherently superior to other architectures, cloud-native principles provide approaches for teams aiming to maximize the benefits of cloud infrastructure.

4.4.1. Microservices Architecture

Microservices architecture is an architectural style that structures an application as a collection of loosely coupled services, which implement business capabilities. The microservices approach is a departure from layered service-oriented architecture and monolithic application architecture. Microservices architectures enable the continuous delivery/deployment of large, complex applications and allow an organization to evolve its technology stack.

The microservices architectural style manifests several properties: Services are highly maintainable and testable, enabling agile development teams to maintain, test, and deploy their respective services independently of other teams. Services are loosely coupled, so that a service can be changed without requiring any change in the other services that it communicates with – there is a clearly defined and stable interface between services. Services are independently deployable. Services can be organized around business capabilities, and responsibility for each capability can be assigned to a small, autonomous team.

4.4.2. Containerization and Orchestration

Microservices architecture coupled with containers such as Docker enables applications to be packaged into lightweight, isolated units that can be easily combined. The queuing system RabbitMQ facilitates communication between microservices in an asynchronous manner [3-5]. In a microservices architecture, the web-based user interface developed with React.js operates as the container's frontend; the frontend requests data by communicating with backend and frontend dynamic link libraries (DLLs) to access the server. Each backend microservice is encapsulated within its own container. The deployment environment for these containerized microservices is Kubernetes (K8s), which provides a declarative approach to application orchestration, enabling users to launch containers based on predefined specification files. By creating services for the pods, HTTP requests can be directed to collections of the same type of pods through cluster IPs. This configuration allows the web UI to call Kubernetes services and access relevant information by interacting with pods hosting the backend microservices.

Cloud-native AI spans several layers of technologies, from infrastructure to application. Achieving AI at scale necessitates a comprehensive architectural design and the seamless integration of AI technologies, with the goal of delivering AI capabilities via a-service and a-platform models. A critical aspect of cloud-native AI is the design of the underlying architecture. Cloud computing encompasses four main service categories: infrastructure as a service (IaaS), platform as a service (PaaS), software as a service (SaaS), and AI as a service (AIaaS). Based on these service types, there are three primary service deployment models: public cloud, private cloud, and hybrid cloud.

4.5. Data Management in AI Ecosystems

Data represents the foundational layer for innovative AI services, ranging from massive foundation-model training to real-time inference and recommendation. A thoughtfully constructed data-management ecosystem should provide efficient capabilities for data search, data transfer, and data scheduling. Optimal placement of data and trained AI models into the appropriate services and regions is a key task of data scheduling that directly impacts application performance and the overall user experience.

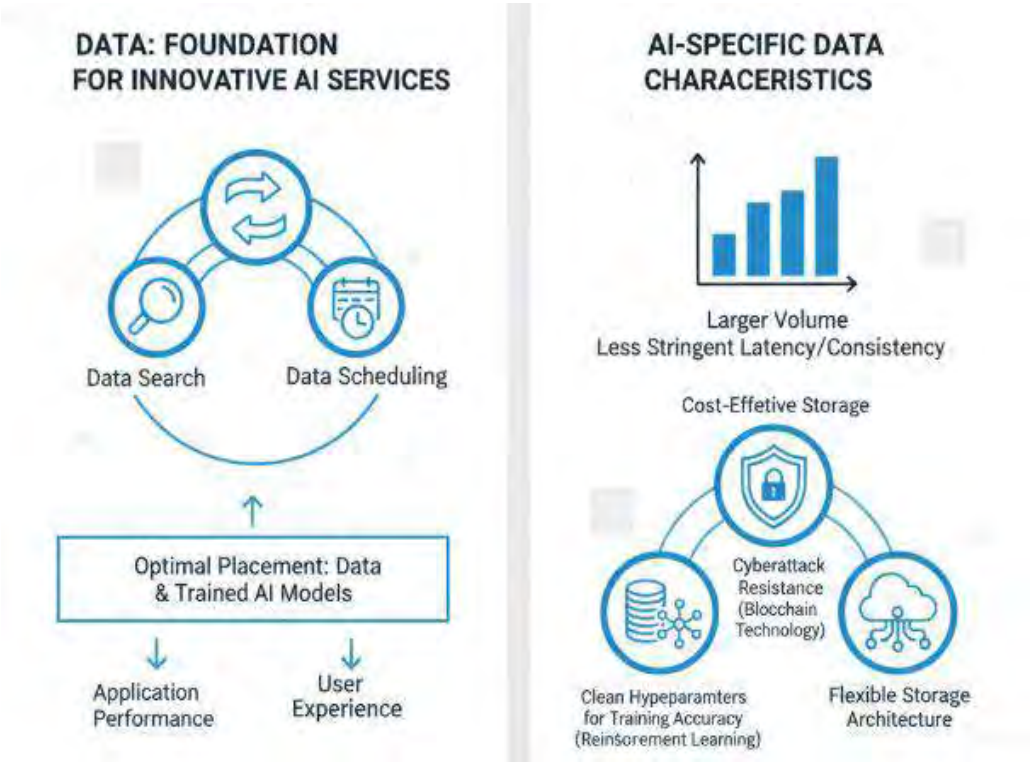

Fig 4 . 3 : Data Management Ecosystem for AI and Cloud Computing

In addition to complex data-access patterns, AI data is often much larger in volume than that for other application types within the same infrastructure, despite typically requiring less stringent latency or consistency demands during access. Addressing AI-specific data characteristics requires a flexible storage architecture that supports cost-effective data management, cleans hyperparameters for training accuracy in reinforcement-learning applications, and resists various cyberattacks with advancements in blockchain technology. Together, AI workloads present a profoundly different set of needs and capabilities for any cloud-computing ecosystem. No sector stands to benefit more from a robust cloud-native design than AI application developers and their users worldwide.

4.5.1. Data Storage Solutions

Cloud-native AI ecosystems are designed to exploit the essential characteristics of cloud computing. These ecosystems reveal the ultimate form of both cloud-native applications and AI development, and are thus nothing but cloud-native AI applications. Their architecture presents a typical cross-section of cloud-native application design. Consequently, the fundamentals of cloud computing and AI introduce foundational elements for constructing cloud-native AI ecosystems.

Organizations leverage AI capabilities through various service approaches, ranging from model sharing and AI algorithm/task sharing to fully trained AI model sharing, deployment sharing, and training sharing. Cloud computing also provides cloud infrastructure services consumed by other AI use cases and applications, such as AI functional services based on analytics or computer vision. Cloud services are divided into three primary categories. AI applications provide business support to end users, ranging from simple robot services to AI devices, employing AI techniques like computer vision, automatic voice recognition, detection, and recommendation. Cloud fundamentals underpin the AI capabilities within cloud infrastructures, encompassing resources management, security, and scalability. Cloud platforms offer self-service solutions, affording users control, development speed, location independence, resource pooling, elasticity, and measured usage [2,4,6]. Cloud platforms or providers own the physical resources and services, delivering service management, service catalog management, demand management, capacity management, and availability management.

The classification system reflects the integration challenges of architecting cloud-native AI ecosystems. Combining cloud-native AI development with cloud-native infrastructure concepts results in a layered architecture model. Data storage follows from cloud fundamentals and design principles, implemented using centralized storage or distributed file systems. Cloud infrastructures group data into databases or cloud buckets.

4.5.2. Data Pipeline Architectures

“What are the different data pipeline architectures used in cloud-native AI ecosystems?” An answer must take into account many important concepts. The discussion begins with fundamentals of cloud computing, together with leading advances in AI. Design principles that arise from the combination then appear, helping to set the context. Next, it becomes possible to move on to the requirements for cloud-native data storage and subsequently to data pipeline architectures. Knowledge of regulations for protecting AI data provides a natural transition to security. Finally, a brief survey of new developments points forward. The answer concludes by emphasizing the challenges of architecting cloud-native AI ecosystems. Cloud computing, provisioned by telecommunication networks, is already well established and represents a paradigm shift in the architectural modelling of generally-available Business Services. AI, in its various forms, makes these Business Services appear smart and capable of conceptualizing like humans. The core design strategy for cloud computing is a set of services operating in isolation, using the best available technologies for their implementation, and communicating using standards-based interfaces. However, the core design strategy for AI hinders the same approach, requiring all support services, like data visualization, data filtering, data labeling, analytics, and model building, to be architected cohesively, maintaining homogeneity, and closely integrated. Still, it is possible to incorporate the supporting AI Services into the Cloud ecosystem, making them cloud-native but AI-focused, reducing the overall time to market, and keeping the necessary flexibility for future breakthroughs. Data pipelines enable the ingestion, processing, and transfer of data via different platform layers in a structured manner [1,3,5] . The architecture must incorporate the communication mechanism between the various cloud-native AI services. An efficient pipeline architecture will address the trade-off between scalability, resilience, and latency. While queue-based pipelines scale well, they tend to sacrifice latency and resilience for scalability. Application of Event Streaming platforms reduces the latency and increases the resiliency, but additional measures have to be implemented to make them scale.

4.6. Security and Compliance in Cloud-Native AI

Security and compliance aspects establish necessary guardrails for development and deployment. These include local and international regulations such as the Health Insurance Portability and Accountability Act (HIPAA) for health information, the General Data Protection Regulation (GDPR) from the European Union, and the California Consumer Privacy Act (CCPA). Protecting intellectual property in the form of sensitive business data and key processes is equally important to secure the business competitive advantage.

From an AI perspective, regulatory exposure begins with training data. The European Union and the USA follow different approaches. With the aforementioned GDPR, data usage should be both purpose- and time-limited—meaning only with the consent of the data originator or with an anonymous use case. If the underlying model is derived from personally identifiable information, rotatory learning can be more beneficial. The AI Act proposed by the European Union goes further by attempting to regulate the use of AI directly. First, all organizations that offer or use AI in the European market shall comply with the provisions of the AI Act. Provisions break down into the European AI High-Risk categorization, where additional risk mitigation contingency and assessment measures are mandatory, such as a risk mitigation system, conformity assessment, and human monitoring. Services considered of unmet risk include governmental UBIs, social scoring, and exploitative practices targeted toward vulnerable groups.

4.6.1. Data Privacy Regulations

Cloud-native AI requires handling sensitive data and can be subject to data privacy regulations such as the EU's General Data Protection Regulation (GDPR). These regulations ensure individuals have control over who accesses their personal information. Sensitive data can be intentionally protected by storing it on the public cloud in an encrypted form that cannot be decrypted. This, however, reduces AI services' ability to process the data and add value[3-5,6]. Other specialties of data being processed—for example, when it is health-related, as in the US Health Insurance Portability and Accountability Act (HIPAA)—can also influence the selection of cloud services to architect complete AI solutions.

Data privacy regulations frequently emphasize overarching principles of care toward consumer and employee data. These directions may not explicitly guide cloud-native AI implementation but strive to reduce consumer data privacy harm. When engineers are building a cloud-native AI ecosystem, incorporating insights from data privacy research is necessary in order to identify potential failure modes and strategize mitigations ought to they occur.

4.6.2. Security Best Practices

Security protects cloud-native AI data and ensures compliance with national policies and regulations. It applies to all aspects of the cloud environment, including data centers, server resources, network communication, hardware devices, operating systems, and SDKs, and throughout the AI data life cycle: data labeling, secure data cleaning, model training, model deployment and serving, scoring, and security monitoring. Security best practices help enterprises establish and optimize their cloud security framework.

The Cloud Security Alliance (CSA) Top Threats to Cloud Computing: Egregious Eleven identifies the main cloud security threats and challenges. The Bureau of Internet Energy and Information of the Ministry of Industry and Information Technology (MIIT) of China released the Internet Data Security Management Measures (IDSM) to regulate the internet data environment. Drawing on these guidelines, the general framework for security risks and solutions in the cloud-native AI ecosystem is outlined.

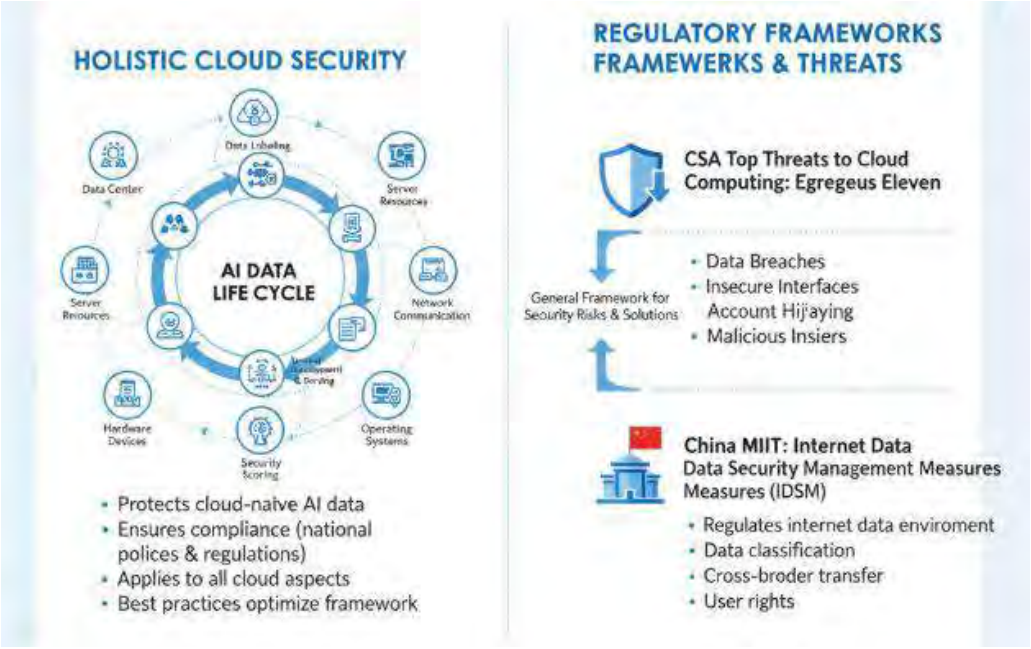

Fig 4 . 4 : Cloud-Native AI Security and Compliance

4.7. Future Trends in Cloud-Native AI

Cloud-native AI generates infinite applications by combining different AI and Cloud technologies. It starts with simple practices using frameworks such as Keras, AI-as-a-service offerings, and public cloud services such as Microsoft Cognitive Services. Services such as chatbots or speech-to-text translation are prototyped, and document classification service chains may be built with the help of Apache NiFi. However, cloud-native AI truly appears with business-driven models, in which AI services support the development of business activities, and technologies such as microservices, containers, and CI/CD are incorporated [6,7,8]. The resulting distributed business models enable the evolution of important Internet companies such as Netflix and Facebook.

The design and development of end-to-end AI services that use complex features from several AI paradigms also belong to this category, for example in a multiturn chatbot in which natural language processing and machine learning work together. Nonetheless,

distributed business models still depend on AI products developed in the centralized model, as exemplified by Google TensorFlow, Facebook PyTorch, or IBM Watson. Another important aspect is the use of containers to provide AI platforms on-premises to satisfy the requirements of data privacy. A final trend in cloud-native AI addresses the growing breadth and depth of services and demonstrates the consolidating influence of Asia in the availability of AI services in the Cloud.

Real-world AI experiences help define best practices that shape the cloud-native AI ecosystem. Devotees of microservices creation comprehend the synergy between microservices and AI as an enabler of the next generation of scientific and business applications with new levels of automation and integration. Data scientists working on the implementation of AI practices further appreciate the benefits that the cloud-native approach brings in terms of productivity, scalability, and an agile framework to support continuous delivery.

4.7.1. Emerging Technologies

Cloud computing technologies are continually evolving, from multi-cloud to hybrid-cloud schemes. In AI, the pursuit of larger and more intelligent models demands ever-expanding GPU clusters. The complexity of these models and the corresponding GPU clusters calls for optimized architectures to reduce training costs, mitigate risk, and minimize resource waste. The emergence of cloud-native AI naturally addresses these challenges.

Multi-Model Serving. In production environments, multiple AI models coexist. Although various serving solutions exist, none are entirely cloud-native. Adopting a cloud-native approach—embedding a model inside a container image for isolated management and serving on Kubernetes in a standardized manner—unlocks significant economy of scale in serving. Moreover, GPU resources, often underutilized, can be repurposed to support online model serving when dedicated GPUs are inactive. The introduction of NVIDIA's Triton Inference Server underscores the importance of multi-model serving. All models follow predefined input–output APIs and can be served instantaneously through RESTful or GRPC services.

4.7.2. Predictions for AI Evolution

The emergence of internet-scale cloud, serverless, container, and container orchestration technologies was met with some criticism in the early 2010s, labeled as hype. However, the current trend in AI has prompted the emergence of cloud-native AI ecosystems, which are conceived and constructed using these architectural features. They offer

several advantages unavailable to traditional AI ecosystem architectures. Looking ahead, improvements in cloud-native AI ecosystems will build on these core capabilities.

While the evolution of AI functions will continue to be driven by innovations in fields such as natural language processing and computer vision, the resolution of challenges associated with the process itself—particularly concerning industrialization, dataset creation, and training—will benefit significantly from cloud-native AI ecosystems. In fact, this perspective distorts the rapid advances made in these disciplines by the research community.

4.8. Conclusion

Cloud-native AI represents a modern approach to AI application development and deployment that exploits the full potential of cloud computing. It enables the efficient harnessing and analysis of massive amounts of data. Cloud computing lends itself well to AI workloads because it provides virtually unlimited storage, near-infinite scalability, free platform upgrades, and minimal maintenance for hosting AI services.

The key design principle for cloud-native AI is leveraging a true cloud architecture model that is inherently scalable, resilient, and manageable. Cloud-native-AI reigns supreme as the only paradigm capable of delivering the compute intensity, scale, cost-effectiveness, and agility required to turn data assets into intelligence. This chapter integrates coverage of cloud computing services, public, private, and hybrid clouds; fundamental technologies used in artificial intelligence; cloud-native architecture design principles; and data management for AI. The last sections survey security and compliance issues for cloud-based AI as well as emerging trends for cloud-native AI.

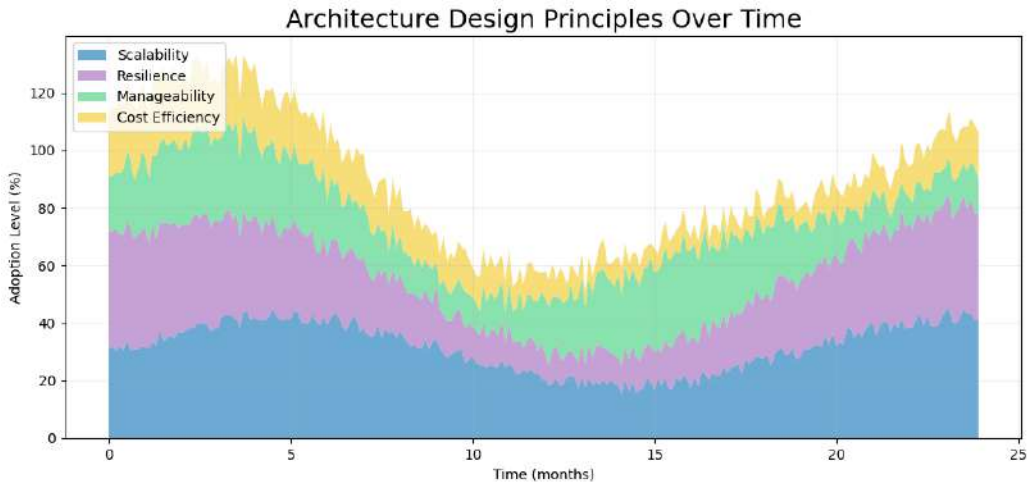

Fig 4 . 5 : Architecture Design Principles Over Time

The main challenge in developing cloud-native AI solutions is the architecture: designing the structure of the system so that AI-related tasks take full advantage of the benefits of cloud computing. Cloud computing hosts the AI platform on the Internet based on a scalable, manageable, and secure cloud architecture. This enables AI architects to develop and deploy AI applications that harness unlimited AI resources and deliver scalable AI services to end-users on-demand. Although the computing power is large, the cost is reasonable. Architecting cloud-native AI ecosystems combines the explosive growth of AI with the powerful capabilities of the cloud, opening a wide range of possibilities in the digital transformation of organizations.

4.8.1. Final Thoughts on the Cloud-Native AI Landscape

Cloud-native AI consists of AI systems architected, built, deployed, and operated on cloud computing infrastructure that fully leverages the benefits of cloud-computing service models and deployment models as well as technologies for cloud-native applications. Architecting cloud-native AI ecosystems presents challenges not only on cloud-computing infrastructure but also on the area of AI algorithms and frameworks. The broader benefits include scalable resource management, sustainable resource utilization, cost-effectiveness with pay-per-use, high availability, minimal maintenance overhead, faster and more powerful innovation, and handling jobs not previously possible.

Cloud computing provides on-demand and elastic computing and storage resources supported with professional operation and maintenance of datacenter-grade infrastructure to protect computing jobs from disasters, utility-grade pricing, and business-grade support. Cloud computing delivers these resources as services with a pay-per-use pricing model and, hence, a cloud computing infrastructure is a utility computing infrastructure, running business applications and providing clouds of services following the natural business growth and slowdown cycle. VM, container, and function-based cloud computing services address part of the challenges faced by AI infrastructures by addressing essential scalability, elastic capacity, and high availability. Designing a complete general-purpose AI system architecture requires more design principles for cloud-native operation and maintenance, as well as full integration with AI technologies and frameworks supporting AI efficiency, model scalability, and application coverage.

References

[1] Zhang J, Tao D. (2020). Empowering Things with Intelligence: A Survey of the Progress, Challenges, and Opportunities in Artificial Intelligence of Things. arXiv preprint.

[2] Pallav Kumar Kaulwar. (2025). Leveraging AI, ML, and Gen AI in Automotive and Financial Services: Data-Driven Approaches to Insurance, Payments, Identity Protection, and Sustainable Innovation. Journal of Information Systems Engineering and Management, 10(36s), 1118–1136.

[3] Siam S I, Ahn H, Liu L, et al. (2024). Artificial Intelligence of Things: A Survey. arXiv preprint.

[4] Munnangi, A. S. M., Nayeem, S. M., Koppolu, P., & Munnangi, S. R. (2025). Experimental and molecular dynamics study of molecular interactions in γ-butyrolactone–dimethyl formamide systems with machine learning based density predictions. The Journal of Chemical Thermodynamics, 107545.

[5] Panduman Y F, Funabiki N, Fajrianti E D, Fang S, Sukaridhoto S. (2024). A Survey of AI Techniques in IoT Applications with Use Case Investigations in the Smart Environmental Monitoring and Analytics in Real-Time IoT Platform. Information, 15(3):153.

[6] AI-Based Financial Advisory Systems: Revolutionizing Personalized Investment Strategies. (2021). International Journal of Engineering and Computer Science, 10(12).

[7] Radanliev P, De Roure D, Van Kleek M, Santos O, Ani U. (2019). Artificial Intelligence in Cyber Physical Systems. arXiv preprint.

[8] Kishore Challa. (2025). AI and Cloud-Driven Transformation in Finance, Insurance, and the Automotive Ecosystem: A Multi-Sectoral Framework for Credit Risk, Mobility Services, and Consumer Protection. Journal of Information Systems Engineering and Management, 10(36s), 1084–1102. https://doi.org/10.52783/jisem.v10i36s.6706