Linguistic and Philosophical Investigations
ISSN: 1841-2394, e-ISSN: 2471-0881
Vol 21 (1), 2022
pp. 1 -15

AI-Native Cloud Platforms: Redefining Scalability and Flexibility in Artificial Intelligence Workflows

1. Srinivas Kalisetty, [2]. Ravi Kumar Vankayalapati, [3]. Lakshminarayana Reddy Kothapalli Sondinti, [4]. Shashikala Valiki,

[1]Integration and AI lead, Miracle Software Systems, srinivas.kalisetty.ic@gmail.com, ORCID: 0009-0006-0874-9616

[2]Sr. software engineer, Equinix Dallas USA, ravikumar.vankayalapti.research@gmail.com, ORCID : 0009-0002-7090-9028

[3]Sr software engineer, US bank, Dallas USA, lakshminarayana.k.s.se@gmail.com, ORCID: 0009-0003-2070-3213

[4]Research Assistant, shashikala.valiki.researcher@gmail.com, ORCID ID: 0009-0008-2853-668X

Abstract

Cloud computing and AI have conventionally been delivered as distinct services. Over the past few years, however, they are increasingly becoming integrated and integral components of a wider service offering. This essay focuses on the intersection of AI and cloud and introduces AI-native cloud platforms—platforms that are developed with AI, big data, advanced analytics, and machine learning as a 'default' integrated part of the platform design.

Founded on the underpinnings of AI as a service and cloud-native AI, this essay argues that AI-native cloud platforms represent the 'future' cloud platforms. We introduce scalability and flexibility as cornerstones of AI-native platforms and show how these platforms deal with key challenges plaguing the current AI platform frameworks. We draw from a case study to showcase real-world cloud-based AI-native platforms. The case example in this essay provides a good theoretical and practical lens for understanding the cloud as AI and portends a future world where AI becomes integrated into, rather than encapsulated from the cloud. Thus, the propositions we present will be of interest to both IS researchers who are interested in cloud and AI as well as to practitioners who may be involved in the design and development of contemporary data platforms and software applications.

With opportunities in the environment surrounding AI-native cloud platforms, this essay seeks to move beyond both the developments and the challenges unique to the cloud as AI. At the same time, we argue that looking at AI as part of cloud computing will enable us to begin addressing some of the contemporary issues in AI that are associated with development in the 'other' direction, i.e., from data and AI models out to real-world applications.

Keywords: AI-native platforms, cloud computing, artificial intelligence, machine learning, big data, advanced analytics, scalability, flexibility, AI as a service, cloud-native AI, platform design, real-world applications, data platforms, integrated services, IS research, practitioner insights, AI integration, future cloud platforms, development challenges, and case study insights.

Received: 08 Nov 2022 **Revised**: 12 Dec 2022 **Accepted:** 25 Dec 2022

1. Introduction

Cloud computing is recognized as a highly sought-after service in AI workflows, in particular as a critical tool for training machine learning models and inference deployment. AI models employed for real-life large-scale applications utilize cloud-based computing services provided by several cloud platforms. As AI and cloud computing are interconnected, advancements in AI are significant in the context of cloud

platforms. One of the main ongoing trends in AI research is to develop domain-specific mathematical formulas. Dominant cloud platforms sponsor these communities, and cloud users interact with community-developed models and tools in diverse ways. Although these developments in AI and cloud computing offer several advantages, there are increasing demands for robustness, scalability, and high functionality of AI workflows for large-scale data streams from diverse areas such as health informatics, financial data stream analytics, autonomous vehicle platforms, and real-time industrial data processing, to name a few.

A survey report outlined that the average percentage of applications using microservices hovers at 65%, with vertical industries such as telecommunications and the healthcare industry attaining the highest levels. Many cloud services have APIs for TensorFlow, as well as for AI-on-Cloud platforms. A recent study has also shown growth in market size for deep learning data processing in the cloud of around 135 billion USD in 2020 to 369 billion USD by 2025. These trends align with the vision of API-led connectivity. Part of the connectivity model emphasizes greater reliance on pre-developed capability functions in the cloud. AI-native cloud platforms are increasingly in demand due to difficulties leveraging recent advances in AI on traditional cloud platforms to enhance performance. Thus, to avoid bottlenecks observed in today's cloud environments and to unleash the true potential of AI technologies, AI-native systems are the next logical step for more flexible and scalable environments.

1.1. Background and Context

Cloud computing technologies have been instrumental in democratizing artificial intelligence for diverse applications, showcasing impressive evolution over the last decade to support increasingly powerful requirements of AI or machine learning workflows. Each year sees the release of more sophisticated hardware and software stacks from vendors to tackle dramatically increased expectations, and we have witnessed the emergence of AI-native cloud platforms as the next logical step. These platforms are not only capable of executing complex machine-learning routines but are also designed to streamline data preparation, facilitate automated machine-learning pipelines, and enable more complex AI operations such as feature engineering, model serving, model management, and optimization. In this context, the coupling of machine learning with cloud computing marks another step in a decades-long wave of silicon-provided capabilities in helping organizations handle the rapidly growing data deluge.

Indeed, several seminal landmarks have traced various evolutions in AI, machine learning, and their convergence with cloud computing. The entrepreneurship of GCP, the commercial release of AWS Lambda, the advent of serverless computing, the growth of cloud services in AI, and the "as-a-service" revolution have helped AI practitioners access, develop, and operationalize increasingly scalable distributed training and prediction solutions offered in the cloud. The urgency for even more scalable solutions continues to be echoed in various works. Over the last few years, our economies have been undergoing a digital transformation with accelerating digitization of virtually every industry. As a result, the cloud computing industry has recently been growing at a swift pace to meet additional demand. These trends make it even more imperative than ever to look at cloud infrastructure technologies that are tailored to the needs of AI and machine learning workflows.

Equation 1 : Dynamic Resource Allocation

$$R(t) = R_{\text{base}} + \Delta R \cdot f(D(t))$$

Where:

$R(t)$: Total resources at t,

R_{base}: Base resources,

ΔR: Additional resources,

$D(t)$: Demand at t,

$f(D(t))$: Scaling function.

1.2. Research Problem and Objectives

1.2 Research Problem and Objectives

Traditional cloud computing platforms are not designed for AI workloads but rather follow paradigms and architectural principles grounded in warehouse-scale computers and utility computing. Reinventing AI-native cloud platforms necessitates the revamping of the storage, computation, and machine learning layers to accommodate the unique characteristics of AI workflows. In AI workloads, it is often the inference step and not the training, which is the compute bottleneck. To gain significant improvements, such traditional cloud computing paradigms must be questioned.

In this research, we seek to investigate AI-native cloud platforms that have been designed to serve AI as a first-class citizen. More specifically, this thesis is generally motivated by the following challenges posed by AI workloads: on one hand, the need for scalability is driven by the need to train machine learning models with orders of magnitude more data over a performance-critical timeline. Each training step requires iterative model execution with vast quantities of data; distributed training accelerates the model convergence, but the full spectrum of data complexity can only be tested molecule by molecule in a shorter period. On the other hand, an insatiable requirement for computing resources, specifically hardware accelerators, demands platforms to be flexible and programmable.

Overall, the primary objectives of this research are to investigate and devise state-of-the-art solutions for these limitations. While a multitude of academic literature supports this research field, literature tends to fall into two categories: theoretical analyses and systematic constructions of AI-native cloud platforms exploring limitations, usually at the level of an edge computing scenario. Nevertheless, to authentically address these challenges in both our discussion and the recommendations made based on it, we find it crucial for the research to encapsulate the continuum from theoretical investigations to empirical and practical results. As such, these research objectives lay the groundwork for both research and industry markets.

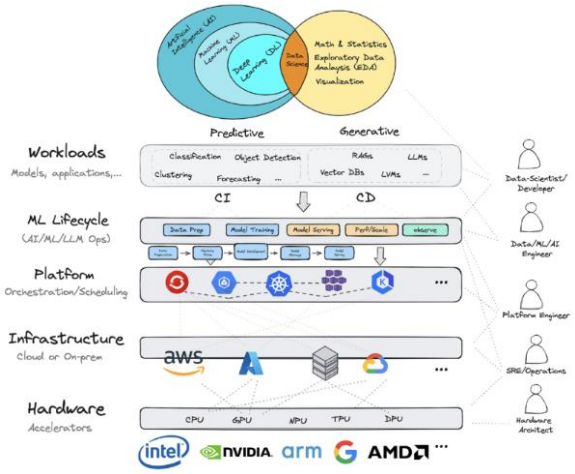

Fig 1 : Cloud-Native AI

1.3. Scope and Significance

1.3 Scope and Significance We assume that AI-native cloud platforms can efficiently manage the convergence of AI algorithms with their data and innovation. The workflows of AI systems are composed of a set of interrelated software components that interact with user data and operate on AI algorithms to provide optimal performance results. We concentrate on those cloud platforms that have the following key aspects to be most relevant to AI workflows, and therefore we expect them to be the most "cloud-compatible" for AI applications: scalability to accommodate the convergence points of AI workflows, flexibility to provide research with multiple tools and algorithm combinations, and performance optimization algorithms to optimize the elastic cloud infrastructure used during AI workflows. Our research focuses mainly on the data-driven, AI-centric big data processing component and external storage. This work thus provides a comprehensive experimentation pipeline integrating task, workflow, cloud

platform, and hypervisor costs and performances. Then, taking into account industry inputs, we discuss the main and novel open issues emerging from experimentation and field deployment and describe some emergent solutions to tackle them. Because AI embraces several areas of technology—such as smart cities, intelligent transportation, and cloud services—these deployments and the solutions are also influencing AI-native research and development as a whole. The results will be illustrated experimentally using the particular case study instantiation on the infrastructure. A scalability experiment will be performed using the public commercial cloud, resulting in a total data production much larger in comparison with previous benchmarks.

2. Evolution of AI and Cloud Computing

Artificial intelligence (AI) technologies have evolved significantly since their inception. AI's development has been characterized as proceeding through multiple rounds of growth punctuated by breakthroughs. Some researchers trace the psychological roots of the field of AI back to ancient times and, more logically, to the 1940s by computer scientist Alan Turing. The early period of AI is characterized by work on reasoning, search, and knowledge representation, typified by the General Problem Solver developed in the 1950s.

The birth of cloud computing in the early 2000s also set the stage for major developments in the field of AI. Before the advent of general-purpose infrastructure, small workgroups within companies tended to build narrow solutions. AI work was sometimes distributed within standing data center infrastructure. Larger applications such as search, customer recommendations, and fraud detection go back to various cloud infrastructures that predate general-purpose infrastructure. The maturation of cloud AI offerings led to the scaling of training workloads. Tighter integration led to one arguably conceptually indistinguishable platform, but little schema or application portability across consumer-centric and enterprise-centric products.

AI advances are leading to a dependence on cloud infrastructure. Scalable model design and training have moved away from the supercomputer and on-premises data center, leading to a push towards traditional map-reduce or other isolated or restricted-access infrastructures. This shift has been exemplified by the increase in AI labs found at large cloud providers over the past 10 years. These labs symbolize not only the solid and perhaps self-sustaining position that cloud providers hold in the technology industry but also what is likely a further consolidation of power around specific cloud platforms.

Fig 2 : AI in cloud computing

2.1. Historical Overview of AI

AI is not a new field but has been the subject of research for over 70 years. The AI domain can be said to have been born in 1956 when researchers convened at a summer school founded by John McCarthy and Marvin Minsky and coined the term "Artificial Intelligence." The historical overview discussed below highlights some important events leading to the development of what we now know as AI.

Early algorithms: The history of AI can be traced back to the very creation of algorithms in mathematics and philosophical reasoning that relate to what is now called AI. For example, algebra and symbolic notation were inspired by the desire to automate what was then seen as the "mechanical" processes of old Greek geometry. The quest for "automating thought" has been a goal seen in later mathematical notation and artifice. There have been many such problems over the centuries that have, in various ways, specified

the subject matter that has concerned AI. This has included puzzles, games, and logical stunts such as those by Rameau in the eighteenth century and others in seventeenth-century England and France. Philosophers such as Descartes, Leibniz, Boole, and Frege were explicitly or implicitly concerned with a description or simulation of thought and its connection to symbolic form and logic. The early methods of mathematical notation and philosophy evolved into modern-day problems in AI, including theorem proving, expert systems, and intelligent agents. In general, however, the tools and conventions of earlier work became relevant only much later.

Behaviorism: Behaviourism is based on the philosophy of "operationalism," which specifies that a concept is equivalent to a specific experiment. Although operationalism has now been largely rejected, it provided one of the first modern formulations of the position. The general tenet and links to AI are given by the idea that the final knowledge of physical objects is gained through mechanistic operations, perceptions, sensations, then thought, elucidated observation statements. The AI equivalent of operationalism is "operationalist semantics."

2.2. Cloud Computing and AI Integration

Cloud Computing and AI Integration

Cloud computing infrastructure has been gaining traction since vendors began delivering computing solutions on a pay-for-what-you-use basis. When regarded in conjunction with AI technologies, cloud infrastructure's inherent nature of scalability and elasticity in data storage and computing resources matches the demands to harness large volumes of data essential for most AI workloads. Furthermore, the cloud's centralized data storage capability makes it easier to connect to other data layers when necessary to enhance decision-making with intuitive business intelligence tools powered by modern analytics and AI, facilitating the convergence of AI technologies. A trend is also observed whereby infrastructure as a service (IaaS) cloud vendors go beyond mere AI infrastructure provision by integrating AI models with cloud services to ease AI model development and deployment via the cloud.

To integrate AI and cloud technologies to allow for this newer kind of workflow whereby AI workloads are deployed through cloud devices to integrate with other data platforms for truly end-to-end solutions, the cloud vendors could embrace AI at different levels of a service, such as:

• IaaS comprising the hardware stack

• PaaS comprising the infrastructure including various core software components

• SaaS comprising entire service offerings. Most cloud vendors already provide human-to-machine models including speech recognition services and chatbots; these can be classified as SaaS where cloud vendors provide the infrastructure and the intelligence to bridge man-to-machine. Successfully developing such an integrated workflow, where a cloud architecture augments AI, would necessitate ensuring:

- Data moving from the edge device (to which the cloud AI model is deployed) must be secure to protect consumer data.

- Any multi-vendor collaboration at the cloud edge must be secure and comply with customer regulations.

Alignment to this new vision suggests that contrary to current theoretical models, a cloud system built for AI requires greater flexibility in the cloud system architecture and utilization. It also necessitates rethinking what is the correct measurement of an AI system in cloud theory. Key to alignment involves moving from a storage/compute-centric system towards defining a system where data starts to converge with compute and AI tools. Several recent high-profile cases signal how cloud-AI-everywhere solutions

are becoming 

a reality.

Fig 3 : Benefits of AI in cloud computing

3. Characteristics of AI-Native Cloud Platforms

AI-native cloud platforms are different in several core characteristics compared with traditional cloud services. Scalability is one of them. AI workloads do not always have uniform computational requirements. In some cases, processing parts of the workload first can be beneficial for the overall end-to-end performance. AI workloads often consist of deep computational parts, followed by structured, fine-grained partial workloads and more deep computation at the end. By separating the computation between GPU and CPU and using multi-stage scheduling acceleration strategies that exploit actual model architectures, they show a performance increase compared to traditional batch-based schedulers.

Flexibility is another critical characteristic of AI-native cloud platforms. As the size and complexity of AI models increase, auxiliary or pre-processing data requirements also rise. Moreover, different users and organizations use different models to train, be they advanced models or models with some simplifications. As a result, an AI-native cloud platform should be able to adjust itself in response to different user demands. When task characteristics and infrastructure are heterogeneous, a cloud-native architecture can make autonomous decisions via heuristics that consider various factors, including workload arrival rate, task duration, and available resources. In addition to choosing optimal strategies for placing tasks, designing a transformation engine heuristic for AI-native ingestion to selectively perform transformations can yield performance gains.

Performance optimization is a critical characteristic of AI-native cloud platforms. The capacity of cloud-native platforms to match resources with task demands can have a significant impact on user experience. For AI-native cloud platforms, several architectural considerations are critical for supporting high-performance AI workloads. For example, AI-native cloud platforms should be able to dynamically allocate resources in response to an increase in the arrival rate of tasks. In most real-life scenarios, an intelligent AI-native platform should be capable of handling such elastic demand rates. AI platforms also need to be able to handle both bare computational characteristics, such as high I/O computation and high memory requirements, as well as varying sizes in data complexity for neural network models.

$$T_{\text{total}} = \sum_{i=1}^{n} \frac{W_i}{C_i} + \delta_{\text{comm}}$$

Equation 2 : Workflow Optimization

Where:

T_{total}: Total time,

W_i: Workload i,

C_i: Compute capacity i,

δ_{comm}: Communication latency.

3.1. Scalability

Scalability is an essential property of AI-native cloud platforms that represents the ability to adapt resources according to changing or growing loads. The AI workloads can greatly vary from small experiments that typically require a few tens of cloud instances to obtain the result within a couple of hours to large-scale deployments with up to multiple thousands of instances. AI-native cloud platforms can

support a variety of workloads by dynamically allocating resources based on the demand. They adapt these at runtime to guarantee the utilization of the cloud infrastructure while being under the constraint to respect service level agreements about the quality of service and volume of data processed. The support for dynamic, rapid scalability is particularly important in AI because it allows trying more ideas faster and radically innovating; this approach is heavily used in AI and is known as evolutionary search.

The nature of workload in AI is bursty, which can differ greatly in volume, duration, and resource requirements. Adaptive, elastic cloud computing is an intuitive strategy to handle it effectively. Scalability is often related to the support for microservices, a design strategy to create loosely coupled components based on business demand. In the cloud environment, microservices are instantiated on demand by allocating the necessary resources automatically. Therefore, resources are adaptable to the current cloud workload and can be used efficiently. Cloud and HPC resources are typically designed to scale, both in terms of throughput and storage capacity as well as the number of simultaneous users. Integration or co-location of AI and HPC results in novel, scalable infrastructures. Several real-world examples are available with scalable cloud infrastructures, which are elastic, adaptive to the workload, and combinatorially efficient.

Despite the benefits, problems can persist in cloud computing, especially in the realm of AI. Scalability can lead to high costs, and therefore sophisticated cloud management techniques are needed to optimize this resource usage effectively. Managing elasticity in efficient ways is an important open problem, especially for AI workloads. Overall, these represent the state-of-the-art technologies to address elastic, scalable solutions, and the main initial requirements of AI-native cloud platforms.

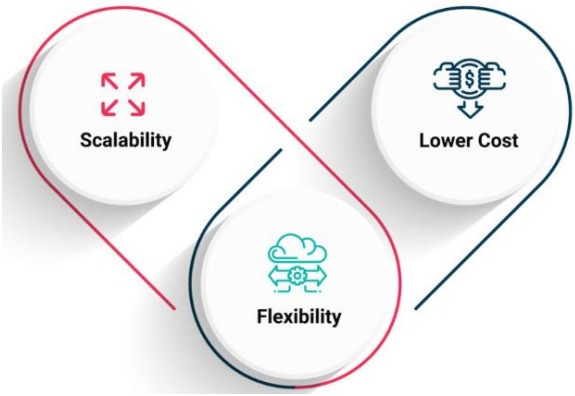

Fig 4 : The Hybrid Cloud Architecture Work

3.2. Flexibility

3.2. Flexibility AI-native cloud platforms are designed with the flexibility required to adapt to a wide variety of AI models and algorithms. There is an extensive variation in the computational requirements of different models and algorithms. The increasingly complex and diverse AI that is being developed to meet specific research and industry-driven challenges means that it is difficult to predict the specific image or signal processing, model learning, architecture, and reasoning techniques that will be employed in the future. Developing solutions that work across a wide range of AI-driven problem areas and allow rapid AI model and framework choice enables a great deal of experimentation and subsequent innovation. It is unlikely that AI capabilities and limitations are fully understood in the science and technology community, so developers need the flexibility to try new ideas and fall back on mature, robust methodologies. Integrating different AI development frameworks and tools is often a complex activity. Specialist computational hardware or libraries, which are tailored to specific AI solution tools, may sometimes seem too different to work together. In the AI-native cloud, it is desirable to have a variety of AI model development and lifecycle tools available. These may use different computational strategies, pattern-matching algorithms, and existing models. An AI-compatible cloud environment makes it easier to test emerging AI technologies in real systems. By providing access to existing and familiar platform components, requirements for testing can be reduced, and the time for testing AI solutions can be estimated. The ability to switch AI solution

deployment in the cloud by changing platforms can enable the rapid deployment of new AI or AI-learned production controls and remove the need to extensively change the entire IT system. A poor performance metric may be improved if the model is rebuilt with some additional processing steps. This may involve changing the data, model, or AI tools used.

3.3. Performance Optimization

Performance optimization is increasingly recognized as a key characteristic of AI-native cloud platforms to deliver cost-efficient, reliable, responsive performance under dynamic workloads. Different techniques can be employed, such as load balancing and resource allocation, among others, to enable IT resources to be used more efficiently and to extract more value from limited resources. Computational efficiency measures how quickly and cost-effectively resources can be optimized for a faster user experience. Meanwhile, system efficiency aims at resource utilization and high throughput, which also drives down costs.

Quantitative metrics can be used to measure the platform's performance, such as cost per user, response time, throughput, and system utilization to satisfaction ratio. The cost per user represents the overall cost of delivering a user request, while user request or job size ranges from several requests to hundreds or thousands. The response, in a probabilistic sense, is the time it takes within a given QoS requirement from the moment a user request arrives until the system returns the user request response. Moreover, the system's effective throughput varies depending on whether it is an interactive or batch/dependence analysis system. System throughput is an essential measure of how many job operations can be done per unit time, and the system response time is the total time a job spends in the system. However, system effectiveness is driven by the response time of individual and dependent job-case analysis, and system effort is the system throughput for the type of job done to support the user community. Technologies for system effort to increase system performance and reliability include containerization, where one would want to balance cost, throughput, flexibility, and speed, and allow for different approaches to resource allocation. New serverless computing could be considered while evaluating trade-offs addressing AI and ML workloads. In addition, infrastructure innovations can ensure high performance, global presence, reliability, and value-added services, such as multi-tier caching, redundancy, encryption, and security.

Lastly, because an AI-native cloud needs extensive mathematical and engineering efforts in infrastructure, systems, and network science, a demonstrated working approach could not only win the attention of the market but also gain market share due to differentiated technology. Despite all the challenges, there have been weeks of affecting resource allocation techniques on satisfying performance metrics for AI and visualization workloads. Customers were satisfied with the results, as clusters used to apply on the order of half the deadline. However, hindrances include minimizing the effects on existing resources to deliver systems and cloud platforms with low user satisfaction. As these infrastructure systems cannot be replaced quickly, minimally viable systems are needed to build up to solve constraints and replication problems. Furthermore, resource allocation, balancing costs, and effort used with tracking budgets could become expensive, and competing products must be considerably competitive.

4. Case Studies and Applications

4.1. Healthcare

Home recovery of post-operative breast cancer patients through the use of video chat and machine learning on the severity of a hematoma. This leads to a better quality of life for the patient and a cost reduction per patient. Use of predictive analytics on the probability of developing eating disorders in well-trained athletes. This is used to better proportion trauma among nurses in healthcare settings. A secure federated machine learning model opens the door to different privacy requirements for similar projects.

4.2. Manufacturing

Prescriptive maintenance, reducing unplanned downtime for machines in the production line of a large multinational. A key pivot point was the realization that the impact of unplanned downtime is limited for customer experience and mostly affects the company's bottom line due to the stockpiling of products ready

for shipping. Instead of focusing on further reducing the already low false positives of the system and further increasing the precision, it became key to enable knowledge-sharing between human operators and the AI in a user-friendly and low-threshold manner. This has not only made the system much better and ultimately allowed for the shipping of disconnected machines, but also provided a base for new industry offerings aiding operators in doing rounds. Model-driven manufacturing. We are moving from AI-driven processes to intelligent manufacturing, where the AI becomes part of the products and their design process. Via generative adversarial networks, we now work on model-driven decisions. This becomes most visible in spare parts production, where we no longer repair, salvage, or otherwise treat a scarce commodity: we redesign the products based on our knowledge of the additive manufacturing processes used.

4.1. Industry Applications

AI-native cloud platforms are capable of being incorporated into applications across a broad spectrum of verticals, driving innovation in a plethora of use cases. Emerging trends in the healthcare and pharmaceutical industry, global finance and retail, public transport agencies, automotive and manufacturing leaders are taking steps to reinvent core functions by leveraging AI-native cloud platforms. Almost anything and everything has AI placed strategically as an enabler of performance, such as operational efficiency, innovative services or solutions, enhancing decision-making, informing recommendations, and improving daily activities. Emerging Trends There is no limit to the potential of AI when it comes to customizing its applications to conform to industry needs. System integrators are leveraging AI-native cloud platforms to move the industry towards new TRP and the development of hybrid solutions that make efficient use of data and existing infrastructure. The most significant market trends in the application of AI-native cloud platforms are related to healthcare analytics, better patient management, data monetization, AI in the banking industry, computer-controlled gaming algorithms, private security solutions, significant reduction of cybercrime due to advanced AI security layers, optimization in appraisal and diagnosis, and smart city trip planners. However, it should be noted that the resultant industry outlook is bullish, although it might be too early for this optimism to take a very strong hold. The implementation of cloud-based AI technologies in the health and financial industry also faces some prevailing difficulties, especially with technological challenges: finding suitable applications and integrating multiple versions of diagnostic aids are significant problems in explaining the results of clinical tests. The security and privacy of data, as already mentioned, and the protection of queued data from being compromised is a major and legitimate concern among decision-makers. Thus, strategies must be in place to face regulatory reform on data privacy and protection, as well as to prepare the medical workforce to change and develop policies, consider solutions to minimize risk, as well as encourage the use of digital platforms. In the banking sector, strategies to manage or respond accurately to the risk of wealth changes affect how a profitable bank can grow, and prominent financial organizations have long employed AI as an enabling technology to collect, store, and leverage data about customers' interactions.

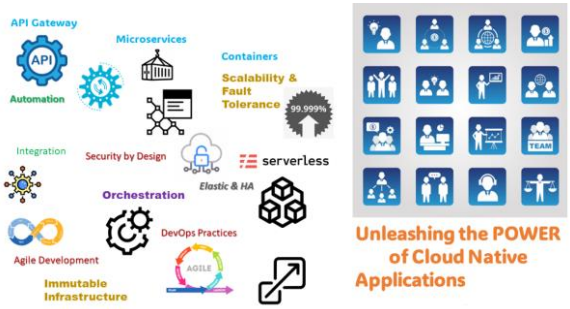

Fig 5 : Unleash the POWER of Cloud Native Applications

4.2. Research and Development Use Cases

Research and Development Use Cases Academic and scientific organizations work with AI-native cloud platforms on leading-edge initiatives for innovation, emphasizing experimentation and discovery. They expose the value of AI, working at the cutting edge of practitioners and facilities to pave new signposts for

AI capabilities and strategies. Large AI models and algorithms could be enabled to reveal scientific discoveries or pave the way for a new research notion that sets the context for AI and computers. Use cases are in support of empirical research in various academic disciplines. The studies might make policy predictions based on real-world deployment into ethical processes and privacy during data training and model deployment. In this case, the cloud allows researchers access to cloud resources by using AI/NLP resources when they need platform scalability and testing. Key cloud differentiators include scalable resources for complex natural language processing with the use of AI, natural languages, and computer vision methodologies, and the ability to change and alter direction through training and edge testing while being able to support different cycles. This example also showcases a resource immediately scalable in terms of one week's traffic volume. As we dig deeper into each use case contribution, learnings and findings will be demonstrated. This is centered around research projects that have passed the IACUC and IRB, or that haven't been required to submit for IACUC review at the time of submission.

5. Challenges and Future Directions

AI and cloud computing present their own unique sets of challenges and opportunities for integration. One of the concerns with cloud computing is the ethical guidelines and regulations affecting where data is stored, and cloud providers may be compelled to host data in certain locations. Further, the use of machine learning and AI requires large amounts of training data that may be subject to license restrictions, particularly concerning personal data regarding individuals based in the European Union. Organizations must also appropriately address AI-related technology ethics and transparency considerations, or they risk the expedited development and deployment of AI algorithms being rejected by users and subject to additional regulatory scrutiny. The importance of lawful, ethical, responsible, and safe AI and resulting ethical and regulatory oversight will need to be maintained to ensure widespread acceptance of AI and the systems it powers by the general public. Any solution to provide cloud platforms as AI processing resources must account for these enabling technologies and provide guidelines on their use based on ethical as well as regulatory standards.

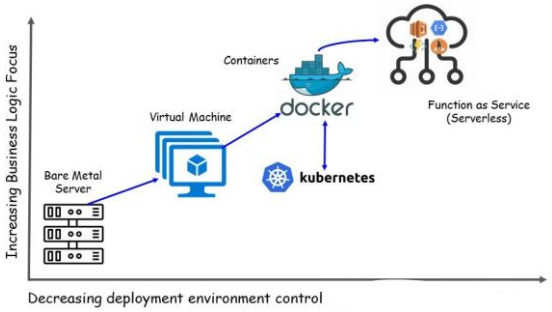

Fig 6 : Cloud native technologies fit together

Duality: A convergence is required. While the management of legal issues regarding culture, humanity, and transparency is a priority, the definition of cloud services powered by AI can be supported by the right to be forgotten and access rollback principles. However, new regulations similar to existing frameworks will be required based on technological innovation and how it may be used responsibly, particularly with services or technologies where the harm or potential for harm is not physical, but still significant — obscured decision-making, privacy, confidentiality, etc. New guidelines will have to be developed around the use of AI for personalization, human intelligence, and interaction. AI-powered systems will have missed the AI ethics and bias scrutiny stage if issues only come to light when a person is notified that someone is no longer working at a company or that they have been disqualified from a job application by an AI tool; in these situations, the algorithms have already been in use, unmonitored and unmanageable. The convergence of the legal governance required for AI, the capabilities of AI in analytical and formulating decisions, and the convergence of managed services customized to AI brings multiple primary and secondary risks to the forefront of AI-powered cloud platforms. This risk brings into focus any enabling technologies or tools that could be used to manage, mitigate, remediate, or minimize the likelihood and

impact of events related to AI and cloud safety. Undertaking this research would well forecast new technological solutions for an AI-native cloud system. It is evident that progression in this area is at an early stage; this indicates that organizations and developers can learn from these challenges and invest in a new cloud market that looks to take advantage of AI-powered negotiation. Taking a proactive stance would help transform any duality into an opportunity, effectively future-proofing safe and mature research for the AI and cloud area.

5.1. Ethical and Regulatory Concerns

The amalgamation of AI and cloud has stirred several ethical debates. The issue of data privacy becomes even more pertinent in the context of cloud computing due to the inherent user-related and non-local nature of IoT sensor data. Other ethical issues, like AI bias, which deals with data discrimination and manipulation, make AI models less ethical by introducing a strong privacy invasion risk. The AI procedures are also required to be more interpretable, which enables their users to check for errors while also ensuring accountability of various stakeholders, notably firms and government institutions, in dealing with AI resources. Given this future possibility, it is important to include AI ethics in the broader scope of cloud computing. Currently, there are no broad-based regulatory frameworks that govern AI, though some sub-areas of AI are undergoing standardization efforts. There is consensus, however, among global researchers and government agencies that AI ethics and AI governance should be included in cloud ecosystems.

The ethics of AI believes that ethical guidelines also need to start from the development stages of AI systems and run through their lifetime. A significant body of recent research investigates and identifies the requirements for ethical AI; frameworks for responsible AI engineering are suggested by many works in this regard. Several recent works discuss the moral implications and violations of a social network service provider in the context of privacy. Once the ethical guidelines are established, regulatory clarification and enforcement are required to follow the responsible development, deployment, and operation of AI-native cloud-based industrial systems. It should be noted that AI systems are complex and they dynamically learn from streaming data while carrying out big data analytics. The development of regulatory frameworks requires careful attention to ensure that innovation and growth opportunities are not stifled. An important area that is still under discussion is the design of administrative and legislative mechanisms to ensure compliance of AI-heavy industrial systems with developed standards. Ensuring compliance is the responsibility of the organizations and the industrial plants that are developing and deploying AI-native cloud-based industrial applications. While AI applications and algorithms are used by organizations, they need to ensure that they comply with the GDPR so that trust in the system can be established. In the case of a violation, the Data Protection Officer should be contacted and an investigation made. Thus, AI-native clouds are a multi-faceted approach requiring innovations in technology, governance, and human infrastructure. To ensure AI compliance with GDPR, trust in cloud platforms is necessary. The regulatory policy should be used with much care; it should not be too narrow nor too broad. Effective legal policies and laws should be made which will lead to fair and effective law enforcement. The governance in the United States envisions a relatively light-touch regulatory policy for innovation. The European Union's governance envisions a more overarching control and direct role in legislative rule drafting. Considering all these points, we believe that AI-native cloud governance should harmonize both sides, striking a balanced approach among various stakeholders. The technology-driven, hands-off, and competitive production are USPs of the United States approach, while the protection of social rights, ensuring justice, and establishing communal harmony is the European Union's approach.

5.2. Technological Advancements and Innovations

Innovations and technological advancements continue to broaden the frontiers of AI-native cloud platforms. Edge computing, designed to move computing and storage capabilities closer to the location where data is generated, unlocks new opportunities and functionalities. Initiatives such as edge intelligence funds and a focus on edge analytics have established how the edge is set to become the next frontier in computing. Similarly, advancements such as federated learning hold the potential to enhance both the performance and functionalities of AI-native cloud platforms. It is a decentralized, secure approach

wherein different devices can work together to learn from decentralized data that remains on the original device. This technique is rapidly being adopted in major industries such as healthcare to advance technologies like AI-based disease prediction, building IoT devices to make predictions about sensor data, intelligence-driven cybersecurity to prevent potential security threats, recommender systems, and applications.

Infrastructural support is critical to providing the appropriate capabilities to an AI-native cloud platform. AI-native workflows are compute-centric, and there is a constant demand for delivering or reducing significant scales of compute resources for them. Therefore, the cloud infrastructure should be able to provide dynamic, on-demand, scalable, and secure environments quite flexibly. AI systems can often be resource-intensive and require the infrastructure to be functional for minutes, hours, or days. Due to this unique requirement, once completed, they can be treated as 'cold resources' to free up data center capacity and optimize costs further. Innovations such as applications and technologies for creating and leveraging AI-native workflows need to surmount many challenges, especially those related to security and privacy, scalability, inclusion of edge resources, distributed/federated computing capabilities, data sharing, soundness, and ethical concerns.

Effectively addressing an ever-evolving technological landscape is beneficial for research, industry, future technological advancements, and the emerging "AI in the cloud" community. Awareness of existing and emerging technologies on the horizon can be invaluable for setting visions, goals, and objectives in the development of futuristic technologies that sustain the current status quo. Further, API-powered AI platforms offer attractive solutions for edge devices. In the HPC domain, the upcoming iteration of Quantum Cloud Services follows a serverless computing model that neatly aligns with AI-centric cloud platforms built on a serverless model. These infrastructural advancements point towards an AI in the cloud ecosystem that is exploring deeper integration with devices on the edge while evolving to cater to the unique needs and requirements of AI workflows.

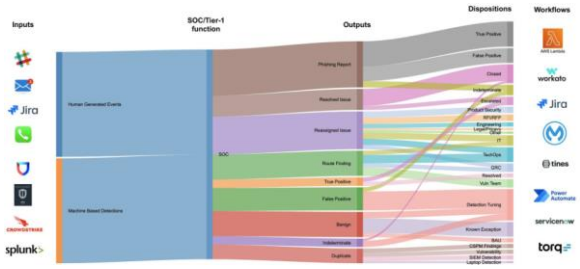

Fig 7 : The Future of SOC Automation Platforms

6. Conclusion

AI-native cloud platforms are an important component in many AI workflows. In this paper, we have discussed their role in developing AI, especially concerning scalability and flexibility. Focuses on moving from AI in the cloud to AI embedded at the edge, and this has a great impact. Scalability and flexibility are not only about making tasks work at a bigger scale, but they enable participants in an AI workflow to experiment with AI and tweak content on the fly. The most important contribution of these AI-native cloud platforms and the convergence of AI and cloud computing for the field of AI is that they enable new experimentation and innovation in machine learning and data analysis and they open AI for use in other fields.

We have seen that AI increasingly is offered as something on top of cloud computing and that it benefits from better infrastructure. One thing that we should not lose out of our focus is that this comes with a price. The price is that AI workflows and software are locked into a specific infrastructure of specific AI-native cloud providers. The challenges of scalability and the lock-in effects can be addressed by increasing the compatibility between AI software and AI infrastructure. Ethical aspects of connecting AI and the Cloud have barely been discussed in this paper. As we see it, both issues should be actively taken up by

stakeholders inside and outside of the field. In general, research on AI-native cloud platforms is still in its infancy – which will continue to grow as the relevance of AI and cloud computing also grows.

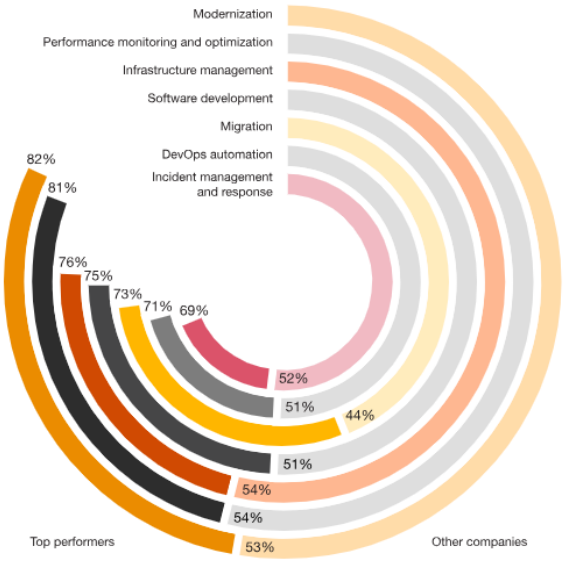

Fig 8 : Cloud and AI Business Survey

6.1. Future Trends

In terms of the technical aspects of AI-central cloud platforms, the future holds several potential paths towards scalability and flexibility. Research may pave the way for technologies such as fully connected systems, more powerful mainframes, more adaptable RTOS, on-chip integration of complex stack architectures, faster serial data networks, or SCIs, WDM for optical computing, and easier access to quantum computing. These developments will significantly boost AI-native platforms in terms of scalability and flexibility. As technology advances, new demands will arise. Users will likely want even faster implementation times, more sophisticated workflow integration, and more integrable systems. Customers will also expect the AI-native cloud provider to significantly enhance AI, likely for facial image recognition and other profiling that the public is only now becoming aware of.

In the AI space, new technologies may change the workflow of one or more possible AI-native cloud vendors. For instance, quantum computing could improve execution time for some types of AI-heavy workloads. Quantum AI could even render present AI hardware incapable of keeping up if such a system is GPU-type matrix-vector based. Conversely, quantum AI could provide the next breakthrough for research, potentially moving AI hardware in an entirely new direction. Therefore, staying current on the AI hardware scene is paramount to intelligence-based searching and utilizing AI hardware that makes sense for a given application. Developments in AI itself, in machine learning for instance, could change the way some workflows are designed as well, possibly leveraging a significant amount of 'on-demand' commercial compute power. We will now enter into the realm of what AI-native systems and developed workflow capabilities could be in the future. Some of these capabilities and possible futures we are only now scratching the surface of. Staying aware of such trends and working within an enterprise that has the potential to change quickly will be important.

Similarly, whole new markets will develop for public-facing applications in AI. The European Union and Canada have already made strides to protect consumer data more than the rest of the world, so it is worth noting that countries and regions with significant control over data may foster entirely new uses for AI and the cloud. These uses may, in turn, promote development in the AI-native cloud world.

$$C_{\text{total}} = \int_{t=0}^{T} \left(C_{\text{compute}}(t) + C_{\text{storage}}(t) \right) dt$$

Equation 3 : Cost Efficiency

Where:

C_{total}: Total cost,

$C_{\text{compute}}(t)$: Compute cost,

$C_{\text{storage}}(t)$: Storage cost.

7. References

[1] Syed, S. (2022). Breaking Barriers: Leveraging Natural Language Processing In Self-Service Bi For Non-Technical Users. Available at SSRN 5032632.

[2] Nampalli, R. C. R. (2022). Neural Networks for Enhancing Rail Safety and Security: Real-Time Monitoring and Incident Prediction. In Journal of Artificial Intelligence and Big Data (Vol. 2, Issue 1, pp. 49–63). Science Publications (SCIPUB). https://doi.org/10.31586/jaibd.2022.1155

[3] Danda, R. R. (2022). Innovations in Agricultural Machinery: Assessing the Impact of Advanced Technologies on Farm Efficiency. In Journal of Artificial Intelligence and Big Data (Vol. 2, Issue 1, pp. 64–83). Science Publications (SCIPUB). https://doi.org/10.31586/jaibd.2022.1156

[4] Rajesh Kumar Malviya , Shakir Syed , Rama Chandra Rao Nampalli , Valiki Dileep. (2022). Genetic Algorithm-Driven Optimization Of Neural Network Architectures For Task-Specific AI Applications. Migration Letters, 19(6), 1091–1102. Retrieved from https://migrationletters.com/index.php/ml/article/view/11417

[5] Patra, G. K., Rajaram, S. K., Boddapati, V. N., Kuraku, C., & Gollangi, H. K. (2022). Advancing Digital Payment Systems: Combining AI, Big Data, and Biometric Authentication for Enhanced Security. International Journal of Engineering and Computer Science, 11(08), 25618–25631. https://doi.org/10.18535/ijecs/v11i08.4698

[6] Syed, S. (2022). Integrating Predictive Analytics Into Manufacturing Finance: A Case Study On Cost Control And Zero-Carbon Goals In Automotive Production. Migration Letters, 19(6), 1078-1090.

[7] Nampalli, R. C. R. (2022). Machine Learning Applications in Fleet Electrification: Optimizing Vehicle Maintenance and Energy Consumption. In Educational Administration: Theory and Practice. Green Publication. https://doi.org/10.53555/kuey.v28i4.8258

[8] Danda, R. R. (2022). Application of Neural Networks in Optimizing Health Outcomes in Medicare Advantage and Supplement Plans. Journal of Artificial Intelligence and Big Data, 2(1), 97–111. Retrieved from https://www.scipublications.com/journal/index.php/jaibd/article/view/1178

[9] Chintale, P., Korada, L., Ranjan, P., & Malviya, R. K. (2019). Adopting Infrastructure as Code (IaC) for Efficient Financial Cloud Management. ISSN: 2096-3246, 51(04).

[10] Chandrakanth Rao Madhavaram, Eswar Prasad Galla, Hemanth Kumar Gollangi, Gagan Kumar Patra, Chandrababu Kuraku, Siddharth Konkimalla, Kiran Polimetla. An analysis of chest x-ray image classification and identification during COVID-19 based on deep learning models. Int J Comput Artif Intell 2022;3(2):86-95. DOI: 10.33545/27076571.2022.v3.i2a.109

[11] Syed, S. (2022). Leveraging Predictive Analytics for Zero-Carbon Emission Vehicles: Manufacturing Practices and Challenges. Journal of Scientific and Engineering Research, 9(10), 97-110.

[12] Rama Chandra Rao Nampalli. (2022). Deep Learning-Based Predictive Models For Rail Signaling And Control Systems: Improving Operational Efficiency And Safety. Migration Letters, 19(6), 1065–1077. Retrieved from https://migrationletters.com/index.php/ml/article/view/11335

[13] Danda, R. R. (2022). Deep Learning Approaches For Cost-Benefit Analysis Of Vision And Dental Coverage In Comprehensive Health Plans. Migration Letters, 19(6), 1103-1118.

[14] Sarisa, M., Boddapati, V. N., Kumar Patra, G., Kuraku, C., & Konkimalla, S. (2022). Deep Learning Approaches To Image Classification: Exploring The Future Of Visual Data Analysis. In Educational Administration: Theory and Practice. Green Publication. https://doi.org/10.53555/kuey.v28i4.7863

[15] Syed, S. (2022). Towards Autonomous Analytics: The Evolution of Self-Service BI Platforms with Machine Learning Integration. Journal of Artificial Intelligence and Big Data, 2(1), 84-96.

[16] Nampalli, R. C. R. (2021). Leveraging AI in Urban Traffic Management: Addressing Congestion and Traffic Flow with Intelligent Systems. In Journal of Artificial Intelligence and Big Data (Vol. 1, Issue 1, pp. 86–99). Science Publications (SCIPUB). https://doi.org/10.31586/jaibd.2021.1151

[17] Ramanakar Reddy Danda. (2022). Telehealth In Medicare Plans: Leveraging AI For Improved Accessibility And Senior Care Quality. Migration Letters, 19(6), 1133–1143. Retrieved from https://migrationletters.com/index.php/ml/article/view/11446

[18] Venkata Nagesh Boddapati, Manikanth Sarisa, Mohit Surender Reddy, Janardhana Rao Sunkara, Shravan Kumar Rajaram, Sanjay Ramdas Bauskar, Kiran Polimetla. Data migration in the cloud database: A review of vendor solutions and challenges . Int J Comput Artif Intell 2022;3(2):96-101. DOI: 10.33545/27076571.2022.v3.i2a.110

[19] Syed, S. (2021). Financial Implications of Predictive Analytics in Vehicle Manufacturing: Insights for Budget Optimization and Resource Allocation. Journal Of Artificial Intelligence And Big Data, 1(1), 111-125.

[20] Syed, S., & Nampalli, R. C. R. (2021). Empowering Users: The Role Of AI In Enhancing Self-Service BI For Data-Driven Decision Making. In Educational Administration: Theory and Practice. Green Publication. https://doi.org/10.53555/kuey.v27i4.8105

[21] Danda, R. R. (2021). Sustainability in Construction: Exploring the Development of Eco-Friendly Equipment. In Journal of Artificial Intelligence and Big Data (Vol. 1, Issue 1, pp. 100–110). Science Publications (SCIPUB). https://doi.org/10.31586/jaibd.2021.1153