*International Journal on Recent and Innovation Trends in Computing and Communication*
*ISSN: 2321-8169 Volume: 10 Issue: 12*
***Article Received:*** *25 July 2022* ***Revised:*** *12 September 2022* ***Accepted:*** *30 November 2022*

# Secure and Automated Cloud Platforms for Next-Gen AI Applications

**Phanish Lakkarasu,**
**Staff Data Engineer, ORCID ID: 0009-0003-6095-7840**

**Abstract**

Since the presentation of ChatGPT by OpenAI in November 2022, the topic of Artificial Intelligence (AI) has been present even among a non-specialist audience. Basic ideas of AI or Machine Learning (ML) were presented to a wider audience on the example of ChatGPT. More and more AI-generated contents in the form of text, images, audio or video are observable on social media platforms. AI supports other apps and online services. Consequently, there is a growing interest in AI and Automation in general. There is probably no such newsletter or magazine which does not have AI in its title, editorial or news flashes. Whatever subject is treated, there is at least one article about generative AI, chat GP, stability AI, ML, etc. However, Machine Learning (ML) has been used for more services in private, commercial and industrial sectors in recent years. Cloud service providers came up with an increasing variety of infrastructure and services that allow cost-efficient and scalable implementation of ML applications. ML predictions can be computed via machine pools with up to thousands of GPUs that are spray-based on demand and the works can be scaled down if they are no longer needed. For many ML applications, cloud services are quite central as they provide a fast, scalable, flexible and cost-effective infrastructure for running sophisticated ML models. With the recent advent of public Large Language Models as a Service, this development was further accelerated. Some key benefits of cloud services for ML are present which have led to an increasing number of ML applications. However, AI and ML applications are not only possible with cloud services. There are areas of application for ML, like autonomous driving, where connectivity to cloud services is not continuously possible or does not make sense.

In autonomous driving, for instance, merging image and radar data on the current traffic situation must occur in real time. It is imperative to have sufficient computing power in the vehicle for processing, computation and decision-making to occur on the spot. Without sufficient computing power on-premise devices, the output of ML models can be obsolete or even misleading. Edge devices are no replacement for cloud services. Cloud services are typically provided by a third-party trust level (or higher). An argument against cloud services in favor of on-premise hardware is better control over data protection and information security. Rather sensitive or company secret data remains in-house where companies are able to better monitor and protect it. Even more so in times of mass data protection violation indiscretion. Private clouds are defined as those cloud infrastructures that are dedicated to a single organization. Fog computing is a term coined to extend cloud computing capabilities to IoT devices and other edge devices. An example of an ML application in fog computing is processing sensor data in a smart grid system. In a typical fog computing architecture, sensor data can be processed at edge devices. ML models can make predictions about energy demand and perform optimizations. Processing data at the edge can reduce latency and improve system efficiency. All said, cloud services are still part of the autonomous driving ecosystem which provides benefits that edge devices cannot offer alone. Users of autonomous driving may rely on cloud services to calibrate on-premise or sensor systems. Personalization is another example, where information about previously observed scenes may be stored in the cloud for better on-premise decisions. Even vehicle-to-cloud communication is conceivable.

The edge computing paradigm is a middle ground between classical centralized cloud computing and the IoT. When computing is closer to the user, the ML result is faster and more accurate. For that reason, this is crucial for applications like AR glasses or autonomous driving where broadly speaking a continuous bandwidth must be provided. An essential prerequisite for correct functioning of AI and ML applications based on cloud services or fog computing networks is reliable cloud services or fog computing

*International Journal on Recent and Innovation Trends in Computing and Communication*
*ISSN: 2321-8169 Volume: 10 Issue: 12*
***Article Received:*** *25 July 2022* ***Revised:*** *12 September 2022* ***Accepted:*** *30 November 2022*

networks. Compromise of cloud services or fog networks will lead to problems and even correct functioning of ML applications. This paper presents security challenges for ML applications based on cloud or fog computing and provides guidance on how to mitigate respective threats.

**Keywords:** Application-Centric Edge-cloud Collaborative Intelligence, Cloud-edge Overlays, Computational Intelligence, Security Challenges, AI Applications, Cloud AI Platform, Data Management Scheme.

## 1. Introduction

In recent years, Machine Learning (ML), especially Deep Learning (DL), has been widely deployed and successfully applied in various domains, demonstrating impressive and even supreme performance in terms of accuracy and latency. To this end, tremendous computation and network resources are required to deal with the rapidly increasing size of ML/DL models and vast amounts of training data. Cloud computing is very attractive to IA (Intelligent Application) developers as the predominant high-performance computing (HPC) paradigm. Cloud providers offer diverse services like Infrastructure-as-a-Service (IaaS), Platform-as-a-Service (PaaS), and Software-as-a-Service (SaaS) to facilitate the implementation of applications. A variety of mainstream IAs are deployed on the Cloud by major Cloud providers to leverage centralized resources for computationally-intensive AI tasks and high scalability of resource acquisition.

However, with the continuous proliferation of modern IAs on the Cloud, novel challenges to CI (Cloud Intelligence) emerge, and have to be tackled when these IAs have to be in production and thus are required to be highly parallel and responsive. For example, IA developers have to keep sensitive data on premises for privacy concerns, and AI inference latency is intolerable when too long network trips to data centers are taken. Therefore, increasing efforts from both academia and industry attempt to exploit heterogeneous resources distributed at the network Edge to address such issues since Edge resources are more close to IA users and have distinct advantages in terms of low latency and personalized intelligent services. Some IAs, e.g. Apple's Face ID and Tiktok [2], offload DL tasks to edge servers for the preservation of privacy in input data and timely response, respectively.

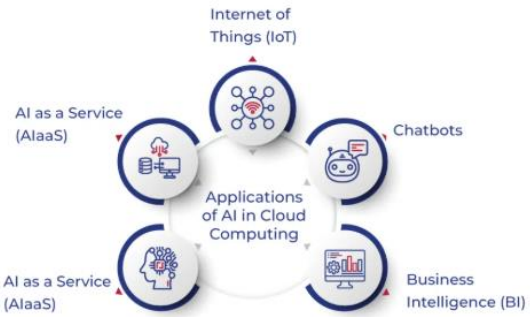


**Fig 1: Applications of AI in Cloud Computing**

### 1.1.Background And Significance

In recent years, cloud-based AI services have gained increasing attention from academia, industries, and end users since they provide convenient and low-cost participation in AI applications from various devices without computation constraints. Cloud platforms facilitating Machine Learning modeling and inference as services have been increasingly adopted. The AI model and data are maintained in a centralized cloud, and clients submit requests to get the prediction results in return. However, as cloud-based AI services proliferate, a surge of privacy issues arises since such AI services require the transfer of sensitive data and models. Adversarial attacks are threats to AI services and have drawn traction. Cloud computing is an emerging paradigm that can virtualize a pool of resources comprising servers, storage, and networks. Users can subscribe to on-demand cloud services to avoid overhead in maintaining infrastructure. Public cloud service providers immensely proliferate public cloud services with reliable performance and security.

A flexible deployment of a pool of cloud resources is also offered with minimal overhead via private cloud solutions. Recently, fog computing has been proposed to extend cloud computing into the edge of the network, establishing inexpensive edge devices as the orchestrators of heterogeneous cloud resources behind the scene. By communicating with telecommunication base stations, fog nodes can interwork with the resource rich public cloud when necessary. Edge devices

*International Journal on Recent and Innovation Trends in Computing and Communication*
*ISSN: 2321-8169 Volume: 10 Issue: 12*
***Article Received:*** *25 July 2022* ***Revised:*** *12 September 2022* ***Accepted:*** *30 November 2022*

can therefore dynamically distribute the workload into multi-layer clouds according to the traffic demand and resource availability, addressing the challenges emerging from edge clouds distributed in a mesh structure. Edge inference is an emerging workload of smart devices that provides real-time and scalable inference services on edge devices powered by low-complexity ML models. It receives increasing attention from academia and industries but still has many challenges in a production system. AI inference services with stringent privacy, latency, and bandwidth requirements are conducted in a tiered manner relying on the orchestration of heterogeneous clouds. AI applications using computer vision, natural language processing, and time-series prediction disciplines are targeted in these systems.

**Equ : 1 AI Platform Efficiency (E):**

$$E = \frac{(C \times A \times R)}{L}$$

Where:

- $C$ = Cloud Compute Power
- $A$ = Automation Level
- $R$ = Resource Optimization
- $L$ = Latency

## 2. Understanding Cloud Platforms

AI has been a hot topic since the presentation of ChatGPT in November 2022. Within a very short time, AI became interesting among a wider audience, not only among experts. There are many discussions on various AI applications, possibly ranging from generating poems or essays to debugging code and code generation or data analysis. In the background of all of these applications, Machine Learning (ML) is increasingly being used for both private and public services in sectors such as finance, healthcare, or agriculture. Due to the required infrastructure and know-how, ML is considered a big hurdle for most enterprises. Developing and implementing (M)L models and algorithms cost a lot of time and effort. Therefore, the fastest trendy way to implement ML projects is to hire and use a cloud service.

Cloud services provide an infrastructure to run ML models and algorithms in a fast, scalable, flexible, and cost-effective manner. Often it is as easy as uploading data and choosing the right setting to get meaningful results and merits from the ML model. Cloud services not only enable ML in businesses but also provide an infrastructure for ML service companies to implement their services. Cloud services are further part of the ecosystem for applications that drive the next generation of autonomous driving. To name only a few examples, traffic flow control based on swarm data from hundreds of connected cars, or map and navigation services dealing with real-time traffic data are just a few of these services. Cloud services are not without any tape. There is an argument against cloud services in any application dealing with important, sensitive, or personal data—better control over data protection and information security. Private clouds provide a dedicated cloud infrastructure only to an organization. Also, fog computing extends the capabilities of cloud computing systems to the cyberspace of Internet of Things (IoT) devices or edge devices.

On the one hand, cloud services offer advanced data analysis, storing capability, and computing power. On the other hand, data protection or information security are still major concerns, paradigm shifts, and strong reasons against cloud services. In contrast to the big cloud services, a third-party provider only provides hardware and basic structures on which the customer has full access and responsibility to manage the IT system. The customer has control over the data, applications, and sensitive information. Therefore, the customers have the opinion that hardly anyone in the organization is able to build such complex systems. Besides cloud services, fog computing can be seen as the opposite of such big clouds.

### 2.1. Types of Cloud Services

In recent years quantitative Data Analytics (DA) have evolved into popular and widely used methods for support of decision-making. Many companies have found that the more aggressively they employ quantitative DA methods, the more successful they become. These examples show emphatically the massive potential and competitive advantage DA brings. In parallel to its success the demand for numerical data of all types has exploded. Sceptics contend that “Data is the new Oil”, because the richest and most successful companies today are “data” companies. Making data available in the Cloud, even at lower cost, enhances its value still further because opportunities for enhanced collaboration, integration or

*International Journal on Recent and Innovation Trends in Computing and Communication*
*ISSN: 2321-8169 Volume: 10 Issue: 12*
***Article Received:*** *25 July 2022* ***Revised:*** *12 September 2022* ***Accepted:*** *30 November 2022*

analysis arise. It is typically complicated to obtain or access data, its format might change over time, and its expectations of quality might not match. To mitigate these issues, all data is co-located and openly available in a collation project. Ideally standardised data storage is provided with a clear API to allow easy access and make it more usable for both informatics and business experts. Over the last years totally new approaches to processing data have evolved (Big Data, NoSQL) to cope with its massiveness and massive incidents of processing. Alternatively to database technology (SQL), graph methods (e.g., RDF) are gaining popularity inside and outside of the scientific data communities. Cloud computing is gaining acceptance as the preferred approach nowadays.

Currently, the industry has been successfully adopting the following common types of cloud computing service models. 1) Infrastructure as a Service (IaaS) is a service model around servers, storage capacity, and network bandwidth; 2) Platform-as-a-Service (PaaS) provides an externally managed platform for building and deploying applications and services; and 3) Software-as-a-Service (SaaS) is having a software system running on a computer that doesn't belong to the customer.

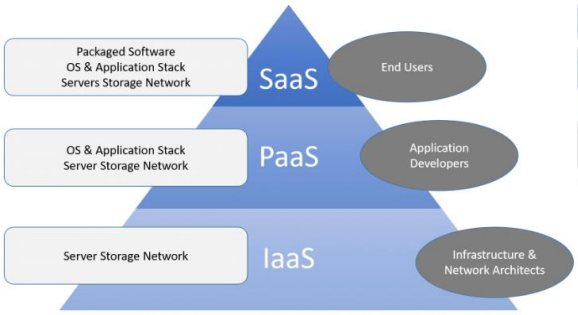


**Fig 2: Types of Cloud Computing Structures**

**2.2. Deployment Models** Cloud-based applications undergo an essential transition from existing on-premise solutions to Software-as-a-Service models. For many ML applications, cloud services are quite central as they provide a fast, scalable, flexible and cost-effective infrastructure for running sophisticated ML models and algorithms. Cloud services allow cloud customers and users to not bother about the underlying infrastructure from which a hosted application might be provided, neither in terms of processing or storage capacity nor with respect to geographic location. Usage of cloud services exclusively through standardised interfaces protects cloud customers from vendor lock-in at least to some degree and allows for switching vendors quite easily. The continuous improvement part of cloud platforms frees customers from (re)investing into and maintaining computer hardware. However, a generalization that AI and ML applications are only possible with cloud services is not permissible. There are also other areas of application for ML, like autonomous driving, in which a connection to cloud services is not continuously possible or does not make sense in large parts. In such areas, the type of application is pivotal, the guarantees regarding latency with respect to processing data points, the probability of false negatives or false positives, and the required availability and reliability of these guarantees, differ considerably from cloud-based applications. This also impacts the type of initial training of ML models and the choice of learning mode (batch learning or online learning).

Edge computing refers to any computer platform, whether physical or virtual, regardless of its size, location, hardware, and software architecture. So, a device such as a microcontroller can also represent a central processing unit in an edge computing context. In edge computing, the processing of data and the execution of applications is generally done on edge devices or devices close to the data sources. This is in contrast to a classic cloud computing model, in which an appliance collects and processes all data in a central instance somewhere in the cloud, potentially far away from the data generation. Still, cloud services are part of the autonomous driving ecosystem, e.g., the higher-level control of traffic flow in a particular region. However, proprietary and privacy-critical data sourced on edge devices does not get sent to the cloud for processing and/or all computation and storage is performed at the edge. Basically, edge devices represent computers that are located between the data sources and the central datacenters. An edge data center typically consists of some servers with possibly attached local storage. Edge computing, with the provision of decentralised complex computing systems such as video analysis, real-time analytics, and real-time optimization of a power-grid, has also risen significantly. However, data production is also rising and the concept of fog computing arose with the aim to extend cloud computing capabilities to IoT devices and other edge devices. The fog layer can help to reduce the latency and bandwidth requirements of cloud computing. An analysis of Internet-connected cameras can be performed on nearby edge and fog devices, detection of objects in images would be performed on the local edge-fog and complex ML models could be

*International Journal on Recent and Innovation Trends in Computing and Communication*
*ISSN: 2321-8169 Volume: 10 Issue: 12*
***Article Received:*** *25 July 2022* ***Revised:*** *12 September 2022* ***Accepted:*** *30 November 2022*

completed by reasoning of detected objects, but only the information that can no longer be inferred from the images would get send to the cloud processing. Furthermore, the scope of the processing can be chosen differently, different fogs could provide different services on the notion of agents processing data on some aspect of the world.

## 3. AI Applications in the Cloud

Cloud computing is increasingly integrating artificial intelligence (AI) capabilities. By doing so, developers can create applications for the cloud that automatically scale to demand, enabling them to serve millions of end users without hitting a bottleneck. On a broader scope, surrounding ecosystems of microservices—presently known as the cloud—enable unparalleled use cases. Such ML applications have the potential to change the way organizations operate, leading to more efficient workflows, lower overhead and labor costs, and new revenue streams. However, the implementation of ML applications also poses substantial challenges. Aside from the efforts required to train models or pipelines and make them available via an API endpoint, successful ML applications should also accommodate robust and fault-tolerant workflows that handle the life cycle of not just the models, but the entire surrounding cloud platform. Such concerns grow more complicated for cloud-based ML, where the underlying infrastructure is managed by a third-party company, which often also remains the party bringing the rock-solid software to the business domain. To address concerns like these, a solution is best built on top of the well-established principles of DevOps and Site Reliability Engineering (SRE) [1]. The first part of the contribution to security challenges for cloud or fog computing-based AI applications is concerned with the security of the training phase and learning data of AI systems. Firms today still deploy, maintain, and train large amounts of cloud or fog systems, but many organizations lack the expertise or tools to protect this vital learning data. Machine learning as a service is publicly available and widely adopted, yet little is known on how to secure such systems and the information they gather and train on.

Automating security systems of provider firms might be an option to enable a solely white-box approach for this phase. However, many sensitive applications, like high-security documents and photo libraries or even targeted advertisements, cannot undergo a black-box approach because their risk has too large consequences. In addition, miscalibrated AI systems might lead to wrong predictions and a failure to fulfill company or organizational goals. The demand for testing such models during development phases grows, yet knowledge of the design of test suites and simulators is likely scarce in provider firms. Employer firms do not want to be aware of the details of learning, but more generally of how the process tests the models in return; as such, the diffusion of knowledge over both systems might quickly provide the firms under attack insight into how to exploit the security loophole.

### 3.1. Current Trends in AI

Artificial intelligence (AI) is a current hot topic, resulting in a high demand for commercial services related to it, especially in regard to the big, general-purpose AI models with exabyte-sized training data and hundreds of billions of parameters. Their training requires enormous resources with data centers of immense size. This creates an industry with tremendous investment opportunities, but also with tough competition. Consequently, AI brings a plethora of opportunities, but also drawbacks: it is increasingly integrated into both everyday applications and specialized areas (medical, finance, etc.) considering both advantages and disadvantages. Unfortunately, the high demand for AI also makes it a highly attractive target for cybercriminals. Since all data and models are stored, maintained, and operated either in the cloud or at the edge, attackers may attempt to poison the models by manipulating the training data. Also, they may try to steal either the training data or the highly valuable model itself. Alternatively or additionally, they may try to prevent access to the AI services on the Internet and extort a ransom from the service provider. By now, it is clear that the rapid development of AI poses a pressing threat to humankind, especially to certain potentially extremely harmful or dangerous applications. However, in contrast to other current big topics discussed and extensively addressed from a security perspective, AI is still in its infancy. The interplay of AI and information security, or the use of AI in regard to security applications, is still an extremely under-explored area with huge potential regarding both applications and research. Also, with the easier and cheaper access to big models and fine-tuners, possible precursors of the major hostile usages may no longer be the sole weapon of a well-resourced evil state. Recently, the first results of using AI against itself have been published. To date, it is already possible to use LLMs as a tool chain to automatically generate tailored phishing emails optimized for a specific target.

*International Journal on Recent and Innovation Trends in Computing and Communication*
*ISSN: 2321-8169 Volume: 10 Issue: 12*
***Article Received:*** *25 July 2022* ***Revised:*** *12 September 2022* ***Accepted:*** *30 November 2022*

**3.2. Use Cases for AI in Cloud** AI is ubiquitous, HYDRA showed next-gen AI multimodal autoML support. Other tools tackle traditional AI tasks, deploying and automating existing models for data ingestion and interaction with users. Tools like CAMEL, openClinica & GNews work on Sensory/NLP data; Tirona automates structured data modeling, Biomechanical augments biomedical, ElectronicLab helps structured tables, and Robocopy helps documentation extraction, but none automate pre-processing & hand-engineering of models. In neuroscience, tools like pNGL and SLAP help non-experts consume existing models. In social science, Rtools do the same for trained classifiers. Workbench and KnowledgeTV analyze co-occurrence networks, and Sentinel works with arXiv papers. However, integration patterns are brittle and depend on platform choices.

Most tools need AI experts for hand-engineering, deployment, and fine-tuning tasks. AI demonstrations on legacy data are showcased, but failure modes hover around impracticality and inaccessibility. Love4Well adds explainability but can't protect health data at ingestion; SOTA pipelines are brittle and require writing ArangoDB schemas and Postgres queries. In Databricks, autoML tools & Hyperopt provide point solutions but lack end-to-end support & credibly entangled ingestion; general providers make safe usage impossible. Huggingface facilitates transformer use but limits interaction slack; multiple-agent interactions are possible, but poorly documented and designed. Late datasets and community transformer training are demonstrated, but generic solutions remain unavailable in research or production. Overall fallbacks have ethical implications; thus, safe usage is limited to codebases owned by AI experts.

**Equ : 2 Security Posture (S):**

$$S = (I + E + M) \times Z$$

**Where:**

- $I$ = Identity & Access Management
- $E$ = Encryption Strength
- $M$ = Monitoring & Logging Capabilities
- $Z$ = Zero-Trust Factor

## 4. Security Challenges in Cloud Computing

Cloud Computing has fundamentally changed today's computational infrastructure for business as well as scientific communities, by providing on-demand and largely-elastic access to a multitude of computing resources. From the resource-constrained edge to ubiquitous Internet-connected data sources, this architecture facilitates the design of new applications leveraging previously inaccessible data and processing resources, while reducing the complexity and the costs of their deployment. However, all of that comes at the cost of data and infrastructure security concerns. This is compounded by the fact that such concerns are growing to an ever-increasing extent, which may completely thwart the realization of the promised advantages of distributed computing for innovating applications targeting business or scientific advancements. As a response to these challenges, researchers point to a paradigm shift towards a new ecosystem, or more precisely an extended cloud ecosystem augmented by security components protecting application services, resources, and infrastructures.

Over the past decade, Cloud Computing has become a dominant paradigm for businesses and scientists to process large scale information and for the deployment of new applications. The idea of offer-on-demand access to computing resources has disrupted established paradigms reducing operational overhead while allowing on-demand scalability of resources. It further promoted a boom of new data- and compute-intensive applications, as data and analysis resources became widely accessible in this ephemeral way. From the data sources perspective, hundreds of millions of Internet-connected sensors collate ever increasing amounts of data every day. Today's data and compute-intensive applications harvest that data from many sources thereby leveraging compute resources at Internet Service Providers or even in the public cloud. Many current real-world applications come to life based on the vision of this Big Data ecosystem. However, the currently available ecosystem contains many problems, which hinders the long term viability of these applications. Chief among them are concerns about the protection and security of data involved in developing, operating, and using these applications. In the current cloud ecosystem, end users, application developers, and service providers are generally not in a trustworthy relationship. Cloud Service Providers (CSPs) are not preventing owned data from falling into unauthorized hands, service providers have almost no visibility anymore

*International Journal on Recent and Innovation Trends in Computing and Communication*
*ISSN: 2321-8169 Volume: 10 Issue: 12*
***Article Received:*** *25 July 2022* ***Revised:*** *12 September 2022* ***Accepted:*** *30 November 2022*

with regard to data used in their applications, hence they cannot be held accountable for misbehaving datasets or results, and it is impossible to assure them of the content and quality of the used data.

Existing and widely adopted auxiliary software tools that strive to provide some of these services are too expensive for small companies and research institutions, need considerable amounts of in-country resources, or are simply proprietary or command and control exploits obscured in black-box services that help neither to build up scientific infrastructure nor to promote competition in the market. To fill the resulting gap between untrustworthy sharing of data and resources and excessive cost and complexity of these services and applications, a new philosophy and a conceptual shift in how computing resources are shared and managed is necessary. The vision is compliance to a framework, consisting of an architecture, protocols, and component designs, relying on transaction-like history-based assurance transactions guaranteeing that security properties hold for a computational process or for the resources allocated to it.

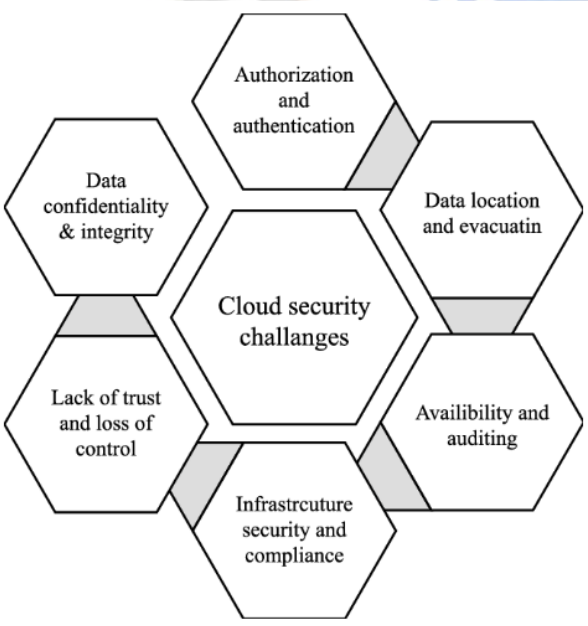


**Fig 3: Security challenges in cloud**

**4.1. Data Breaches and Vulnerabilities**

Based on today's knowledge and experience, there is a general agreement that AI applications either would not work at all or only provide severely degraded services, without cloud self-adaptation and reconfiguration. This is a vulnerability and may provide room for attacks with dire consequences. There are several reasons why this is so. First, data breaches. In applications such as chatbots, personal assistants, recommender engines, and fraud detection, success depends on having massive amounts of high-quality training datasets. This is also true for many multimodal perception applications such as autonomous driving, drone footage analysis, and ground surveillance. This motivation also affects the most damage-intensive applications, as large amounts of valuable data can cause serious harm. Unauthorized access to streaming data or archived datasets needs to be prevented, if only to shield intelligence on to-be-deployed data protection measures. As with conventional data breaches, attackers may try to obtain access through exploiting the cloud infrastructure's surface and lookup for virtual machine images that contain master data for datasets. This requires a certain cloud knowledge. But once this hurdle has been overcome, those types of attacks follow the same paradigm as exploitation of vulnerabilities in well-programmed applications. This is true for cloud services containing Speech/Language Technology or similar solutions. A well-researched approach will reveal the network topology of the cloud application, and future attacks focus on obtaining keys, model weights, or cloud caches hosting model abilities and data used for border fixing repeatedly. Known methods for protecting recurrent and customized speech recognition models are currently missing for vast pre-trained models hosted on distributed processor architecture. Furthermore, exploitation of these vulnerabilities is likely to create pressure to pay for data back as well as release abuse enabling future attacks. A first measure that should be considered is data integrity checking that also accounts for model drift. Cloud-side collaboration becomes crucial as recovered perturbed data requires training compatibility with models. Quality assurance in the context of data quality assessment becomes possible when capturing models themselves. To raise suspicion on failure of model data leaks downstream end-to-end models with online learning deployment need to be retrained. While these measures affect the integrity of available knowledge and successfully mitigate seldom attacks, they also provide feedback for tactics against conventional cloud service architectures. Another vector of attack against AI applications semantically and with high stakes are models trained with a fraction of data volume or conceptually smaller parameter sets but designed similar enough to reproduce original results. Knowledge spreadability of highly modular cloud deployments advocating confidentiality of a few blocks can substantially reduce application performance.

*International Journal on Recent and Innovation Trends in Computing and Communication*
*ISSN: 2321-8169 Volume: 10 Issue: 12*
***Article Received:*** *25 July 2022* ***Revised:*** *12 September 2022* ***Accepted:*** *30 November 2022*

### 4.2. Compliance and Regulatory Issues

Despite the potential societal benefits of large generative AI systems, they pose considerable risks that need to be mitigated. However, a mechanism to enforce compliance with these rules does not yet exist. Generative AI systems require immense computational resources and are thus operated on cloud infrastructure owned by compute providers (CPs), which could intervene to restrict abusive usage of these systems.

This text endeavors to describe some of the technical means by which CPs could verify that customer workloads comply with respective guidelines and regulations. Additionally, the opportunities, challenges, and likely impacts of implementing such measures are elaborated upon. The focus is mainly on large language models (LLMs), but risks linked to other recent prominent technologies largely overlap. Starting with corresponding risks, next, some options for intervention are discussed. Then, a technical account of the kinds of information that would enable such interventions is given. This is followed by a discussion of the governance-related challenges and impacts of enacting such measures.

Abusive large language models can be used for the amplification and spread of misinformation and disinformation, impersonating individuals or organizations, generating abusive content, and creating deep fakes. Such activities can harm democracy and public safety and amplify inequality, bias, and hate. The nature and scale of the risks posed by such LLMs are commensurate with social media and similar forms of technologies increase calls for regulation. Nevertheless, attempts to create suitable legal frameworks are still in early development, and regulations with global coverage do not currently exist. While companies are voluntarily enacting voluntary standards and the EU is working on a comprehensive proposal, compliance controls for these standards are lacking.

## 5. Automating Cloud Security

There has been a growing interest in the development of cloud-based machine learning services and more recently in fog-based services. Cloud computing allows companies and organizations to take advantage of a specified computational capacity without having to invest in the required hardware themselves. A relatively new trend on the edge of the network is fog computing, where the services are hosted on relatively low-computation power devices. The most popular applications for cloud-based machine learning services currently are web services like. These services, mostly intended for company use, allow for the development of reliable machine learning models without needing extensive knowledge of the technical details in their implementation. However, this ease of use is traded in for a potential privacy leak of the data being processed, as it is all transferred to the cloud. Besides transferring data privacy documentation these services can take care of the required building blocks needed for a ML-service like the necessary hardware, data storage, optimization of batches used for processing a trained model, or regularly updating a trained model with new data. In this research, preparation is undertaken in order to identify the security challenges for cloud or fog-based machine learning services. This will ensure that subsequent work concerns contributions to the improvement of security on important and addressed issues. Security challenges for cloud or fog-based machine learning services pose several concerns. First, an overview of the security challenges posed against machine learning services hosted in the cloud is provided. For these services, securing the underlying cloud or fog services is essential, as successful attacks against these services can lead to significant impairments of these applications. Then, the challenges posed on the protection of machine learning services in fog systems are outlined. In this case, the security requirements are different, as these services are hosted on shared hardware devices that are owned by the provider of the fog network, and therefore physical access to them is unobstructed. However, the differences in requirements also open up new avenues for innovation.


**Fig 4: Cloud Security Automation Framework**

### 5.1. Security Automation Tools

Internet security issues such as sensitive data breaches, communicational threats, and fraud are of great concern in our

*International Journal on Recent and Innovation Trends in Computing and Communication*
*ISSN: 2321-8169 Volume: 10 Issue: 12*
***Article Received:*** *25 July 2022* ***Revised:*** *12 September 2022* ***Accepted:*** *30 November 2022*

increasingly digitized world. This growing concern to protect sensitive data on the World Wide Web drove an exodus of data to private cloud servers. The emergence of the cloud brought up several hard problems such as malicious cloud service providers (CSP) colluding with dishonest users or spurious malicious data uploads. Automated cloud compliance analysis on the behalf of the cloud stakeholders is hardly addressed. Addressing the agnostic cloud threat model of honest-but-curious CSPs, several Algebraic Method (AM) based approaches are proposed to allow users to determine whether their outsourced data or computation is kept intact through a definitive aggregation of offline analysis, typically report generation and audit of these reports. However, users can only determine whether an offending violation is committed after they have been accessed and analysed which temple a steganographic approach is followed. Addressing agnostic threat of data breaches, a combination of AM and deterministic encryption based approaches is proposed to protect the query history of compliant users. The presented approaches assume associations of queries and results are uninteresting released, effectively a deterministic ranking of the results. CloudSafe is proposed to understand several important security threats faced in CSPs and a tool named CloudSafe to automate the detection of these security violations. The automated detection of patch processes for either environment is investigated through the development of various policy-driven detection tools. A summarization language that can express such monitoring policy description in a concise manner is provided. A policy-driven and highly portable tool for UNIX/Linux OSes is developed based on that language. In addition, an efficient and general framework of policy-driven detection for post-OS patching processes is designed. The first commercially available tool for file change detection is evaluated and reconstructed. Modifications on to improve efficiency and compatibility with other UNIX OSes are addressed. has been installed and evaluated on a number of real-world systems and a prototype of is built in debugger. Security challenges in cloud or fog-based computing Security challenges for cloud or fog-based machine learning services are outlined. Securing the underlying cloud or fog services is essential, as successful attacks against these services can lead to significant impairments of machine learning applications. The responsibilities for security can be divided between different parties: The cloud providers offer the cloud services, but it is the users' responsibility to store their data securely and restrict access to trained models. Security deficiencies at a lower level can have a direct impact on the higher level where user data is stored. Security deficiencies at these lower services directly affect user data. Responsibilities are simpler for fog computing networks, but services at the edge of the network must be secured against physical access to the devices.

### 5.2. Best Practices for Automation

Automation is an essential part of cloud services beyond the provision of computation and storage. Cloud platforms need to offer services for AutoML, risk prevention, and understanding the usage of resources and costs by analyzing their distribution, as well as visualizing the prediction results of AI models. The next step for cloud services is to completely automate the use of cloud-based AI applications, otherwise known as fully autonomous cloud services. They not only provide computation and storage similar to conventional cloud platforms but also become self-healing and self-evolving so that they can run and improve themselves without human intervention. Be able to automatically obtain computing hardware, automatically propose a model, and automatically learn to optimize themselves. With the increasing number and scale of AI services adopted by multiple domains including smart cities, healthcare, information and education, finance and business, and energy, supporting infrastructure provisioning for AI applications has also become more complex. Various hardware resources such as CPUs, GPUs, FPGAs, and TPUs should be provisioned to meet diverse needs for different learning tasks, and non-AI tasks, IoT connection, storage, and data preprocessing and postprocessing need to be supported as well.

To provide proper resources for each service automatically, cloud platforms generally need to solve a decision problem, for example, using a rule-based algorithm, or manually assessing the previous logs to determine the upper bound for the resources provisioned. First, trained models for these services should be collected, consolidated, and stored. Preprocessing and features extraction in the data lake should also be enhanced to reduce data loads. Then, at the scheduling and/or provisioning point, the resource needs for each service task should be properly forecasted, or at least appropriately assessed. Next, the AI services should be effectively and efficiently positioned based on their needs and influences. Finally, the infrastructure should be elastically provisioned, deployed, and incrementally monitored.

*International Journal on Recent and Innovation Trends in Computing and Communication*
*ISSN: 2321-8169 Volume: 10 Issue: 12*
***Article Received:*** *25 July 2022* ***Revised:*** *12 September 2022* ***Accepted:*** *30 November 2022*

**Equ : 3 Automation Index (AI):**

$$AI = \frac{(P + D + O)}{M}$$

Where:

- $P$ = Pipeline Orchestration
- $D$ = Deployment Automation
- $O$ = Observability Integration
- $M$ = Manual Interventions

**6. Infrastructure as Code (IaC)**

Domain-specific security requirements can be expressed in Security Policy Language (SPL). The security policies will be translated into explicit configuration augmenting the initial configuration. The augmentation will also be validated in terms of security requirement satisfaction with respect to the security policies. The expression will be completely language independent. The valid confinations will be used to beautify the environment in which security is improved. New automated agents can be developed for the new infrastructure as part of the production-ready project evolution. The aesthetic capabilities of the GUI can also be developed in terms of beautifying the environment for existing infrastructure. The automation of the existing infrastructure and the beautifications of the environment will be done via cloud automation platform using Infrastructure As Code (IaC). In addition to existing training, header and indexing customizations can be added to the agents using Domain-Specific Language (DSL). Some capacity customizations can also be achieved using the DSL augmenting the applications. In addition some aesthetic customizations can also be simulated along with the production-ready agents. The cloud automation platform will be augmented using REST API to support the new aesthetics customizations as vague and high level requests will be expressed. Cloud automation platforms using market-leading tools will be implemented. The knowledge base of these automation platforms will be populated using ready-to-use building blocks for a range of existing cloud services. The implementation will also involve the development of Web-GUI and a set of REST API to communicate with the cloud automation platform. The automated deployment will only be partially dependent, i.e. the selection of a few building blocks or a few related configurations will be needed to initiate the whole automated run. Cloud platforms process data at massive scale involving long computation chains and work confront a complex challenge to deploy, test and implement these pipelines. The pipelines include cloud resources creation such as Is (Instances), Containers with CPUs or GPUs, storage initiation such as S3, Bigtable and RDS creation. The pipeline creation will be essentially manual and tedious for existing services with common gain of knowledge and resources from the parent cloud instances. In addition updating and maintaining the cloud services to align with the constantly evolving cloud APIs will also be demanding and cumbersome.

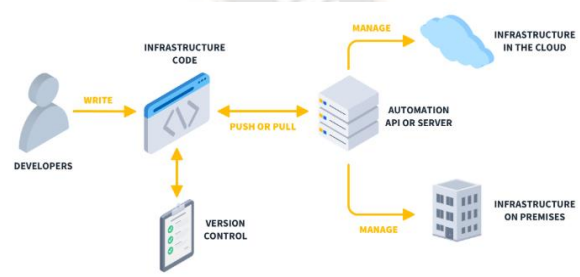


**Fig 5: Infrastructure as Code (IaC)**

**6.1. Overview of IaC** Infrastructure as Code (IaC) is a ubiquitous solution for automating the management of resources provisioned on cloud and on-premises Computing and Network platforms. The management of cloud resources is generally referred to as Cloud Resource Management (CRM) in the context of Cloud Computing and for its usage of Cloud-centric platforms or dominant technologies. As a methodology, IaC allows the specification of the target infrastructure using a domain-specific language, often as scripting languages. These low-level specifications are interpreted by a runtime system that instantiates the infrastructure by sending Management Actions to the Computing and Network platforms used to provision the resources. The Runtime System makes the ideas of Infrastructure as Code (IaC) concrete and applicable. IaC provides a formal specification of the implementation of Management Actions, which are patterns of automation scripts that realize the overall Architecture Task Actions. This separation of specification and implementation of the syntax in Infrastructure as Code allows the latter to be abstracted away by various Modeling Techniques.

*International Journal on Recent and Innovation Trends in Computing and Communication*
*ISSN: 2321-8169 Volume: 10 Issue: 12*
***Article Received:*** *25 July 2022* ***Revised:*** *12 September 2022* ***Accepted:*** *30 November 2022*

The IaC specification can be partially directly generated by Modeling Techniques and/or by the user, while the implementation and its instantiation can be performed by the Runtime System. Although there are numerous off-the-shelf tools in Cloud Computing and Network Management domains, none of them fully comply with the encapsulated architecture proposed above. Some tools cover all layers as monolithic systems, while others provide only low-level management tools. Existing tools focusing solely on the modeling aspects usually lack the capability of automatic code generation. One example of a minimal integration of a modeling tool with a runtime system is provided in. But this has been done for a specific tool, which is neither reusable nor extensible.

### 6.2. Benefits of IaC in AI Applications

Infrastructure as code (IaC) is the management of the entire infrastructure through code and is a key factor in the automation of the Machine Learning life cycle. Intended for unifying the deployment of resources on cloud providers while lowering the level of obtained abstraction, IaC offers a variety of tools to deploy cloud resources, and the most popular ones are Terraform, Pulumi, and AWS CloudFormation. In addition, IaC aims at modifying and deploying resources in a declarative manner instead of manually deploying each component and the entire infrastructure. The deployment of cloud resources is performed via scripts describing the desired specifications of the resources.

A variety of plugins to configure the resources offer a more declarative specification of their properties, leaving the orchestration of the infrastructure to the IaC tools. Resource deployments can be changed either through modifying the configuration files or by modifying the cloud provider's settings. With the assistance of the IaC's tool-driven architecture, setting up or demolishing whole infrastructures would be equivalent to code execution, allowing access to the configurations used for the deployment. Such easy access to the configurations permits versioning and testing of the infrastructure. By employing the API of the cloud provider, IaC provides exact records of the configurations used for the deployment. Similarly, stored in code repositories, these configuration files will provide easy access to a downscaled version of the production environment, permitting detailed experimentation with models or data pipelines. IaC provides a wide variety of tools enabling the abstraction of deploying cloud resources while providing versioning capabilities. As IaC lives along the entire lifecycle of the applications, it will be a key factor in pushing the deployment of MLOps to production-grade infrastructures.

## 7. DevOps and AI Integration

The penetration of Artificial Intelligence (AI) technologies across all domains of human activity is in an exponential growth phase. AI technologies have been mostly used by experts to create new services and manage the extraction of value from large amounts of data. Recently, they are moving towards more generic solutions for various vertical domains and industries. Decisions on maintenance of roads or bridges or how to optimize public lighting in smart cities are increasingly informed by AI models. To increase overall productivity, the adoption by non-experts is supported by efforts to lower the entry barrier by democratizing AI through intuitive systems such as AI as a Service (AIaaS).

Existing cloud service providers developed computing and storage functionalities that enabled an ecosystem for accelerated application development. Additional layers of abstraction and functionality led to XaaS that enable advanced application development and value creation. Various commercial solutions offer user-friendly AIaaS solutions, but their business models rely on controlling parts of the AIaaS technology stack. Recent offerings enable on-premise infrastructure deployment of edge cloud computing where physical resources may be on-premises. Fully controlled on-premises AIaaS stack is challenging and costly to deploy. Therefore, the AIaaS stacks consist of two components: the infrastructure and the model management functionality.

Hiring cloud infrastructure management professional experts may support the development of an AIaaS stack, but limits the democratization of AI. The model management functionality has an overarching goal to minimize the effort necessary to provision models in production, which entails automating the broadest possible set of model management tasks.

*International Journal on Recent and Innovation Trends in Computing and Communication*
*ISSN: 2321-8169 Volume: 10 Issue: 12*
***Article Received:*** *25 July 2022* ***Revised:*** *12 September 2022* ***Accepted:*** *30 November 2022*

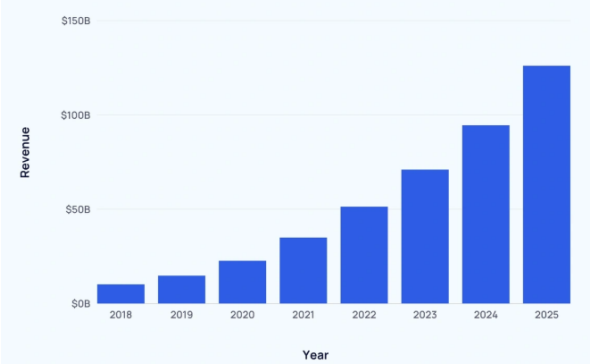


**Fig : Generative AI Statistics**

**7.1. Continuous Integration and Delivery**
Continuous Integration (CI) is a *software build* and testing scheme that encourages individual developers to integrate their code changes into a shared repository many times per day. Each check-in is then verified by an automated build (including test) to detect integration errors as quickly as possible. The goals of CI include increasing the team's awareness of changes to the code or the environment that may affect the functioning of their software; encouraging the team to find and fix defects while they are fresh in mind and involve fewer programmers; and enabling team members to maintain a coherent view of the status of their software in its environment. CI aims at making integrations of software changes more reliable and evergreen. CI can also involve unit tests, automated build scripts, additional tests to replace or augment human testing, deployment scripts and procedures, and configuration management scripts and practices. Continuous Integration is, however, not limited to the domain of software engineering but has been increasingly adopted across domains, including game programming, biomedical data analysis, social sciences, and high-performance computing as well as in the context of Continuous Experimentation. Continuous Integration is claimed to be applicable to these other domains, but it was not obvious which specific practices should be implemented to bring an effect similar to that of CI in the software engineering domain. In addition, more insight into Continuous Integration's applicability across domains may facilitate exploring and performing task analysis for other domains. Continuous Integration practices are claimed to be a key enabler of effective knowledge management in software engineering. Continuous integration allows unit testing, integration, build generation, and deployment to be submitted through a trigger mechanism, rather than relying on manual processes, to run in the cloud using dedicated environments provisioned on-demand to be used only for the duration of the tests.

**7.2. Collaboration between Teams** Big data science will benefit greatly from collaborative efforts among scientists across disciplines. Newly emerging collaborative platforms will permit diverse teams of scientists to pool large data resources and/or run complex combination simulations, bridging multiple physical and temporal domains on heterogeneous computing infrastructure. Open-source backends will support this collaborative effort. Software-controlled cloud-based structures for automatic orchestration of compute resource management will yield rich, heterogeneous computing pipelines. These backends will be progressively outfitted with intelligent user-facing assets that assist the scientist in composing and parametrizing cross-disciplinary simulation combinations collaboratively, visualizing input/output assets and building knowledge databases, and performing large-scale query prediction training. Novel user interfaces employing gamification and inquiry in novel hardware interactions will further allow scientists to push the frontiers of knowledge generation and application.

Artificial intelligence (AI) is currently being integrated into modelling and simulation workflows to directly interact with simulation results during the workflow execution. The incorporation of AI into simulation workflows has two major benefits. First, it significantly reduces the volume of data that needs to be processed, as only simulations with favorable conditions will ultimately be run, thus maximizing resource utilization efficiency. It also precludes the creation of a steady stream of data, as such data requires post-processing before it can be stored. Similarly, as user needs and available data change, adapting all simulation workers to fit a new procedure may be quite costly. Embedding AI into workflows allows the addition of new data and/or procedures that may or may not be related to existing ones without unravelling the overall structure of the workflow. AI is used to train query and prediction models that force simulations to run only in the regions that can enhance or update their results.

## 8. Data Management in Cloud AI

The increasing growth of AI models, the amount of data used to train them, and their complex architectures have led to a high

*International Journal on Recent and Innovation Trends in Computing and Communication*
*ISSN: 2321-8169 Volume: 10 Issue: 12*
***Article Received:*** *25 July 2022* ***Revised:*** *12 September 2022* ***Accepted:*** *30 November 2022*

demand for cloud resources across the entire service stack. These models require more and more data, transfer costs, and exclusive use of the cloud for training. AI operations are conducted on fast-moving cloud technology, causing gaps between cloud vendor capabilities and user growth. Challenges arise from the competitive AI cloud landscape: large and asymmetric information about sender and receiver capabilities and costs, data transfer service latency based on data positioning vs. pipeline processing cost. An approach to AI job service negotiations over multiple clouds is being pursued. Typically, large scale cloud AI jobs are composed of the entire service stack of data storing, preparing, training, and validating. Cloud jobs are expressed as pipelines of processing tasks. Each service creates new information by applying a processing function to its in-memory data. Users of a platform create service job requests containing the information to process. Users, either explicitly or implicitly, convey tentative information to the platform or its operators. The platform has its own questioning policies, cost information, and optimization objectives. It must process information needs without revealing costs or structures. Data transfer involves data transfers between two clouds and advertisements monetizing data. A service monetization mechanism allows the exchange of datasets and advertisement prices. Asymmetric infrastructure costs lead to a possible optimal strategy of deferring additional cloud uses.

There are various tasks for the cloud platform, such as preserving laborious computation, providing disperse job scheduling, strengthening collaborative cloud intelligence, and more interpretable task completion. To offload a subset of tasks from one or more data owners to the cloud, service selection, scheduling, load balancing, location selection, execution environment selection, and result verification need to be addressed. Each task consists of application, input, and output components that must be properly selected and coordinated. One optimization goal is the overall monetary cost for all tasks, which can be further decomposed into sourcing, scheduling, and computation cost objectives. Cloud-related costs, such as data transfer and geo-location costs, need to be detailed for offloading tasks and utilizing cloud services. Complex users, such as companies and organizations with many devices but few individual agents who prioritize profit, account, bandwidth, and pluggability, are also provided.

**8.1. Data Storage Solutions** Cloud Storage model involves storing and managing data and information in a cloud service provider's data center. The data stored with the cloud storage service provider is accessible from anywhere and with any computing device connected to the internet. Cloud storage services can be classified into two types: Storage-as-a-Service (StaaS) and Managed storage services.

Cloud StaaS is a utility-based service consisting of storage and computation as fundamental resources. Numerous service providers offer and manage wide-ranging storage resources that clients can access on demand. Cloud resources including storage are provisioned using application programming interface (API) calls. The pricing for using cloud services is based on a pay-as-you-go manner. Similar to popular computation services that execute big data jobs using cloud resources, some new tools enable users to execute big data jobs directly on cloud storage without needing assigned data processing resources. With the growing data growth, high volume and variety data are now stored with cloud storage providers. Some studies also incorporate cloud storage with some scientific computation pipelines. However, there have been very few works focusing on cloud storage-aware cloud resource usage for data storage optimization in big data analytic pipelines.

Managed Big Data storage services assist users with storing big data by providing a web-based interface to manage storage resources. Users upload datasets through the interface, and the service creates and manages the storage resources in an underlying cloud storage facility. Thus, the data management overhead is transferred to the service provider. The managed big data storage service users either pay a monthly fee of underlying resource usage or a fraction of their job submission charges in a pay-as-you-go manner similar to other cloud services. Big data using managed storage services also avoids data transfer overhead from external storage to the local compute facilities. The decision to use managed storage services is cost-effective since they only involve job submission charges. Nevertheless, it relies heavily on the cloud storage provider. If the cloud storage provider doesn't ensure the retrieved data's consistency and readiness before executing the big data job, the job may fail.

*International Journal on Recent and Innovation Trends in Computing and Communication*
*ISSN: 2321-8169 Volume: 10 Issue: 12*
***Article Received:*** *25 July 2022* ***Revised:*** *12 September 2022* ***Accepted:*** *30 November 2022*

**8.2. Data Governance and Ethics** In the cloud, AI Governance done on a meta level requires appropriate measures to govern the software and hardware architecting a service or solution, its data storage, processing and training, and metrics for regulation and compliance. In a cloud center of excellence (CCoE) AI Governance can be located at the enterprise level with a proper AIMS governance framework adapted to the existing cloud governance framework. AI Services Governance covers the entire data pipeline, i.e., data storing, data processing, data labeling and training, with appropriate service governance measures. Two Governance umbrellas: AI Service Governance Groups (SGs) for data and model as a service (DaaS and MaaS) where the governance focuses on the use of off-the-shelf AI services (AaaS) of a cloud provider; and AI Solutions Governance Groups (SGs) for M/S/A early in development encompass governance for the design-in-figuring (DI-figuring) of models in use, where proprietary data pipelines are developed. Several governance mechanisms regarding broader ethical concerns remain general in rules, such as bias in AI.

XAI in the AI service solutions is highlighted and have been handled in a case study regarding accountability checks of AI for automated grading of student essays. Data Governance measures adapt the existing comprehensive data governance framework. The regulatory aspect of comprehensive data governance has matured in the past couple of decades, but the compliance aspect lags, especially in regard to ethical governance. Data may hold bias under important protected attributes, and regulations do not cover secondary data usage for this purpose. AI service designs are over-parameterized, and AI software experts are needed to extract data, explain which type of data was used for training, and assess whether the software behaviour has been adversely changed. AI Practitioners must thus assess the output of grades and remain accountable in educating and informing end-users, which is often outside of their regulatory scope.

## 9. Performance Optimization Strategies

With the rapid increase in amount and complexity of data, the methods used to analyze data and extract value change in a constant and rapid way. This talk regards machine learning and deep learning as a prominent next-generation application class. A fundamental principle is to extract as much value from the data as possible by applying the best methods. There is a trend to use advanced methods, such as neural networks and deep learning. Well-designed algorithms generally increase resource requirements while improving computation precision. Deep learning is a group of methods with the greatest rise of popularity, and the largest impact, in recent years. A compute and resource intensive method has become a ubiquitous human activity. New deep learning workloads are similar to big data and SQL-based workloads. Whole infrastructure becomes complex and difficult to operate with multiple platforms and technologies. As deep learning and machine learning workloads become dominant, AI-driven cloud services become increasingly popular. In this domain, a platform-as-a-service reachable from the web has the capability to control diverse and heterogeneous cloud services, similar to the web browser. In addition, lower cloud tone becomes the default cloud business model that raises challenges to performance and cost. A large amount of learning resources need to be provisioned. Data, model, and algorithm complexity continue to grow. Workloads become difficult to predict. It may take 1640 hours of continuous compute time to train an image competitor at the current state-of-the-art level. Such workloads are a new class of elephant workloads. The other challenge arises from the wide variety of current cloud services from dozens of cloud providers in public clouds to thousands of data centers in private clouds. It is difficult to choose cloud resources for machine learning and deep learning workloads.

**9.1. Scalability Considerations** The proposed "Hyper" framework was able to harness the capabilities of modern cloud platforms, such as low-cost computation resources, automated provisioning, and on-demand storage to design a scalable Cloud Data Processing (CDP) framework for Deep Learning (DL) based workflows. With the help of cloud technology, there is a far lower barrier-to-entry of parallelism for exploitation of modern DL frameworks. Furthermore, the resources are not only beneficial for speed considerations but also essential to train the very large deep architectures and process the high-volume real-time data. However, the problem of managing fully automated resource provisioning and fault-tolerant in an efficient manner has grown rapidly.

Despite advancements in the specific components of their workload, a well-devised architecture is still missing that provides a one-stop solution and can efficiently distribute the whole End-to-End DL pipelines over hundreds of nodes with

*International Journal on Recent and Innovation Trends in Computing and Communication*
*ISSN: 2321-8169 Volume: 10 Issue: 12*
***Article Received:*** *25 July 2022* ***Revised:*** *12 September 2022* ***Accepted:*** *30 November 2022*

minimal user intervention, fault-tolerance, and efficient communication. Consequently, cloud providers nowadays offer a variety of services which are rather loosely-coupled and require extensive engineering efforts to build a pipeline from them. Numerous technologies have become available as ready-to-deploy services, which are not optimized for DL computations and for the federated manner it is difficult or slow to configure. Thus, the existence of their system exposes such batch DL processing technologies via APIs in the cloud, which can be customized to build complex DL applications on massive data.

The first component is a general-purpose scheduler utilizing cloud function services that wraps the execution of batch jobs over cloud resources. Its serverless nature can automatically scale with the volume of the workload and consequently has very low operational costs compared to other proposals. To host the orchestrator, a mix of utility and high-level cloud solutions are utilized. The cloud orchestration itself is implemented in Beam and it distributes the end-to-end jobs over a decentralized Dask cluster managed by Kubernetes and runs on a low-cost cloud compute service using Linux containers. Furthermore, most end users need not provide cloud credentials as they use the service through API keys.

**9.2. Load Balancing Techniques** Load balancing is a healthy distribution of workloads over a set of resources, with the goal of optimizing resource use, maximizing throughput, minimizing response time, and avoiding overload. The resources may include computer clusters, network links, central processing units, disk drives, or other resources in the cloud and achieving a balanced, optimal distribution is a challenging task. Load balancing as part of resource management is a critical research area in cloud computing with a goal of achieving efficient allocation of resources (nodes) to workload (tasks), utilizing the available resources and ensuring that no single resource is overloaded.

Cloud computing moves computing and storage from local or on-premise locations to remotely hosted servers (clouds). Most cloud platforms provide Infrastructure as a Service (IaaS), where a client has access to a pool of virtual machines (VMs) for processing requests. There is a huge cost to constructing and maintaining such data centers but this cost is offset by the benefits of maintenance and accessibility of cloud vendor knowledge. Cloud computing has the manner of providing services as computing resources like software, network, and storage over the web for end-users that ultimately want to implement a minimum number of systems and hardware to meet their final goal.

The cloud acts as a coordinator and offers services, processing, and storage computing. It makes intelligent resource allocation and task scheduling a challenging issue. Load balancing issues confront with distribution and managing the loads across multiple resources based on their availability and utilizations. When load balancer distributes the load on the resources dynamically, it's known as dynamic load balancers. When it is known before processing, it's referred to as a static load balancer. Hybrid load balancer is a combination of two or more methods to take advantage of techniques. Depending on the types of the resources and tasks, algorithms or a combination of algorithms give effective results.

## 10. Conclusion

Cloud computing has become an indispensable part of an enterprise's IT infrastructure. The cloud services market is anticipated to reach USD 1675 billion by 2030. As organizations expand their reliance on cloud platforms for critical business functions, improving the security architecture for cloud platforms is essential as malicious attacks can lead to significant damage. Current malicious attacks against cloud services often come from collusion among cloud service customers, with cloud service vulnerabilities serving as the initial attack vector. Attack vectors include exploiting infrastructure software bugs, poorly implemented services, and cross-account APIs. Proof-of-concept experiments show how low-level service exploitation can lead to account-to-account (A2A) attacks, impacting billions of user accounts across multiple services.

With multiple collaborative services on a single cloud platform, the integrity of one service may affect the integrity of all collaborative services. Current countermeasures only focus individually on service capacity abuse or service integrity abuse. An automated algorithm is proposed to model multi-cloud service burgeoning cooperative scenarios using directed graphs with heuristic strategies to control an attack path's availability. Additionally, a new cloud production architecture with isolated pipelines is designed and applied to a well-known e-commerce ecosystem with extensive experimental analysis.

This architecture contains on-demand isolated service pipelines that prevent new production from affecting legacy production, allowing for maximum flexibility and minimal cost.

Critical cloud services have become essential parts of the IT infrastructure of various organizations. These cloud services provide a wide range of critical functions from storage, e-commerce, social contact, to social budgeting. The cloud services market is anticipated to reach almost USD 1675 billion by the year 2030. Cloud service providers are responsible for protecting and servicing customers' data before providing any cloud services. The cloud computing security architecture ensures a security-compliant cloud infrastructure for subsequent cloud services. A security-compliant cloud is a well-structured cloud that rigorously satisfies all security requirements.

**10.1. Future Trends** The global economy is experiencing a rapid transformation fuelled by Artificial Intelligence (AI) and the Cloud. With its cognitive learning capabilities and computational agility, interest in AI computing is expanding beyond video surveillance, computational biology, and cancer screening to emerging domains such as self-driving vehicles, financial risk management, customer experience, and smart cities. Recent advances in Neural Networks and DeepLearning algorithms and frameworks that ship with GPUs enhance the ability to extract correlations from increasingly larger datasets in the cloud and at the edge. Furthermore, enormous computational power or clouds of GPU farms are made available for developers, as well as automatic scaling and pay-per-use pricing. This opens up AI usage regardless of capital costs but creates huge infrastructures. As the AI community amplifies and grows, significant engineering challenges emerge in cloud architectures, data supply chains, data integration, and model distributions that raise the operational expenditures.

The success of AI is due to the CHRDL technology stack that is making the massive shift possible in a HaaS delivery model with cloud GPU capacities, design tools, and technologies being made available. Converging technologies such as FPGAs, ASICs, and PhDs are attractive for large deployment but less relevant for the startup ecosystem, which consists mainly of small companies and academic research labs. To predict and visualize the required infrastructures for future AI applications, an analysis of cloud usage trends, distribution of edge-cloud balances, and a forecast on the needs of prospective AI models are needed. The forthcoming evolution of cloud systems will be low-cost Many-Task Computing systems based on considerably higher-density processor units and networks in the AI and ML domain. This will greatly enhance prediction capacities and make AI affordable for small companies and countries, transforming the tech base of the global economy as a whole.

**References**

[1] Kommaragiri, V. B., Preethish Nanan, B., Annapareddy, V. N., Gadi, A. L., & Kalisetty, S. (2022). Emerging Technologies in Smart Computing, Sustainable Energy, and Next-Generation Mobility: Enhancing Digital Infrastructure, Secure Networks, and Intelligent Manufacturing. Venkata Narasareddy and Gadi, Anil Lokesh and Kalisetty, Srinivas.

[2] Pamisetty, V., Dodda, A., Singireddy, J., & Challa, K. (2022). Optimizing Digital Finance and Regulatory Systems Through Intelligent Automation, Secure Data Architectures, and Advanced Analytical Technologies. Jeevani and Challa, Kishore, Optimizing Digital Finance and Regulatory Systems Through Intelligent Automation, Secure Data Architectures, and Advanced Analytical Technologies (December 10, 2022).

[3] Paleti, S. (2022). The Role of Artificial Intelligence in Strengthening Risk Compliance and Driving Financial Innovation in Banking. International Journal of Science and Research (IJSR), 11(12), 1424–1440. https://doi.org/10.21275/sr22123165037

[4] Komaragiri, V. B. (2022). Expanding Telecom Network Range using Intelligent Routing and Cloud-Enabled Infrastructure. International Journal of Scientific Research and Modern Technology, 120–137. https://doi.org/10.38124/ijsrmt.v1i12.490

[5] Pamisetty, A., Sriram, H. K., Malempati, M., Challa, S. R., & Mashetty, S. (2022). AI-Driven Optimization of Intelligent Supply Chains and Payment Systems: Enhancing Security, Tax Compliance, and Audit Efficiency in Financial Operations. Tax Compliance, and Audit Efficiency in Financial Operations (December 15, 2022).

[6] Mashetty, S. (2022). Innovations In Mortgage-Backed Security Analytics: A Patent-Based

Technology Review. Kurdish Studies. https://doi.org/10.53555/ks.v10i2.3826

[7] Kurdish Studies. (n.d.). Green Publication. https://doi.org/10.53555/ks.v10i2.3785

[8] Motamary, S. (2022). Enabling Zero-Touch Operations in Telecom: The Convergence of Agentic AI and Advanced DevOps for OSS/BSS Ecosystems. Kurdish Studies. https://doi.org/10.53555/ks.v10i2.3833

[9] Kannan, S. (2022). AI-Powered Agricultural Equipment: Enhancing Precision Farming Through Big Data and Cloud Computing. Available at SSRN 5244931.

[10] Suura, S. R. (2022). Advancing Reproductive and Organ Health Management through cell-free DNA Testing and Machine Learning. International Journal of Scientific Research and Modern Technology, 43–58. https://doi.org/10.38124/ijsrmt.v1i12.454

[11] Nuka, S. T., Annapareddy, V. N., Koppolu, H. K. R., & Kannan, S. (2021). Advancements in Smart Medical and Industrial Devices: Enhancing Efficiency and Connectivity with High-Speed Telecom Networks. Open Journal of Medical Sciences, 1(1), 55-72.

[12] Meda, R. (2022). Integrating IoT and Big Data Analytics for Smart Paint Manufacturing Facilities. Kurdish Studies. https://doi.org/10.53555/ks.v10i2.3842

[13] Annapareddy, V. N., Preethish Nanan, B., Kommaragiri, V. B., Gadi, A. L., & Kalisetty, S. (2022). Emerging Technologies in Smart Computing, Sustainable Energy, and Next-Generation Mobility: Enhancing Digital Infrastructure, Secure Networks, and Intelligent Manufacturing. Venkata Bhardwaj and Gadi, Anil Lokesh and Kalisetty, Srinivas, Emerging Technologies in Smart Computing, Sustainable Energy, and Next-Generation Mobility: Enhancing Digital Infrastructure, Secure Networks, and Intelligent Manufacturing (December 15, 2022).

[14] Phanish Lakkarasu. (2022). AI-Driven Data Engineering: Automating Data Quality, Lineage, And Transformation In Cloud-Scale Platforms. Migration Letters, 19(S8), 2046–2068. Retrieved from https://migrationletters.com/index.php/ml/article/view/11875

[15] Kaulwar, P. K. (2022). Securing The Neural Ledger: Deep Learning Approaches For Fraud Detection And Data Integrity In Tax Advisory Systems. Migration Letters, 19, 1987-2008.

[16] Malempati, M. (2022). Transforming Payment Ecosystems Through The Synergy Of Artificial Intelligence, Big Data Technologies, And Predictive Financial Modeling. Big Data Technologies, And Predictive Financial Modeling (November 07, 2022).

[17] Recharla, M., & Chitta, S. (2022). Cloud-Based Data Integration and Machine Learning Applications in Biopharmaceutical Supply Chain Optimization.

[18] Lahari Pandiri. (2022). Advanced Umbrella Insurance Risk Aggregation Using Machine Learning. Migration Letters, 19(S8), 2069–2083. Retrieved from https://migrationletters.com/index.php/ml/article/view/11881

[19] Paleti, S., Burugulla, J. K. R., Pandiri, L., Pamisetty, V., & Challa, K. (2022). Optimizing Digital Payment Ecosystems: Ai-Enabled Risk Management, Regulatory Compliance, And Innovation In Financial Services. Regulatory Compliance, And Innovation In Financial Services (June 15, 2022).

[20] Singireddy, J. (2022). Leveraging Artificial Intelligence and Machine Learning for Enhancing Automated Financial Advisory Systems: A Study on AIDriven Personalized Financial Planning and Credit Monitoring. Mathematical Statistician and Engineering Applications, 71 (4), 16711–16728.

[21] Paleti, S., Singireddy, J., Dodda, A., Burugulla, J. K. R., & Challa, K. (2021). Innovative Financial Technologies: Strengthening Compliance, Secure Transactions, and Intelligent Advisory Systems Through AI-Driven Automation and Scalable Data Architectures. Secure Transactions, and Intelligent Advisory Systems Through AI-Driven Automation and Scalable Data Architectures (December 27, 2021).

[22] Sriram, H. K. (2022). Integrating generative AI into financial reporting systems for automated insights and decision support. Available at SSRN 5232395.

[23] Koppolu, H. K. R. (2021). Leveraging 5G Services for Next-Generation Telecom and Media Innovation. International Journal of Scientific Research and Modern Technology, 89–106. https://doi.org/10.38124/ijsrmt.v1i12.472

[24] End-to-End Traceability and Defect Prediction in Automotive Production Using Blockchain and

Machine Learning. (2022). International Journal of Engineering and Computer Science, 11(12), 25711-25732. https://doi.org/10.18535/ijecs.v11i12.4746

[25] Chaitran Chakilam. (2022). AI-Driven Insights In Disease Prediction And Prevention: The Role Of Cloud Computing In Scalable Healthcare Delivery. Migration Letters, 19(S8), 2105–2123. Retrieved from https://migrationletters.com/index.php/ml/article/view/11883

[26] Sriram, H. K., ADUSUPALLI, B., & Malempati, M. (2021). Revolutionizing Risk Assessment and Financial Ecosystems with Smart Automation, Secure Digital Solutions, and Advanced Analytical Frameworks.

[27] Avinash Pamisetty. (2021). A comparative study of cloud platforms for scalable infrastructure in food distribution supply chains. Journal of International Crisis and Risk Communication Research , 68–86. Retrieved from https://jicrcr.com/index.php/jicrcr/article/view/2980

[28] Gadi, A. L., Kannan, S., Nanan, B. P., Komaragiri, V. B., & Singireddy, S. (2021). Advanced Computational Technologies in Vehicle Production, Digital Connectivity, and Sustainable Transportation: Innovations in Intelligent Systems, Eco-Friendly Manufacturing, and Financial Optimization. Universal Journal of Finance and Economics, 1(1), 87-100.

[29] Dodda, A. (2022). The Role of Generative AI in Enhancing Customer Experience and Risk Management in Credit Card Services. International Journal of Scientific Research and Modern Technology, 138–154. https://doi.org/10.38124/ijsrmt.v1i12.491

[30] Gadi, A. L. (2022). Connected Financial Services in the Automotive Industry: AI-Powered Risk Assessment and Fraud Prevention. Journal of International Crisis and Risk Communication Research, 11-28.

[31] Pamisetty, A. (2022). A Comparative Study of AWS, Azure, and GCP for Scalable Big Data Solutions in Wholesale Product Distribution. International Journal of Scientific Research and Modern Technology, 71–88. https://doi.org/10.38124/ijsrmt.v1i12.466

[32] Adusupalli, B. (2021). Multi-Agent Advisory Networks: Redefining Insurance Consulting with Collaborative Agentic AI Systems. Journal of International Crisis and Risk Communication Research, 45-67.

[33] Dwaraka Nath Kummari. (2022). Iot-Enabled Additive Manufacturing: Improving Prototyping Speed And Customization In The Automotive Sector . Migration Letters, 19(S8), 2084–2104. Retrieved from https://migrationletters.com/index.php/ml/article/view/11882

[34] Data-Driven Strategies for Optimizing Customer Journeys Across Telecom and Healthcare Industries. (2021). International Journal of Engineering and Computer Science, 10(12), 25552-25571. https://doi.org/10.18535/ijecs.v10i12.4662

[35] Adusupalli, B., Singireddy, S., Sriram, H. K., Kaulwar, P. K., & Malempati, M. (2021). Revolutionizing Risk Assessment and Financial Ecosystems with Smart Automation, Secure Digital Solutions, and Advanced Analytical Frameworks. Universal Journal of Finance and Economics, 1(1), 101-122.

[36] AI-Based Financial Advisory Systems: Revolutionizing Personalized Investment Strategies. (2021). International Journal of Engineering and Computer Science, 10(12). https://doi.org/10.18535/ijecs.v10i12.4655

[37] Karthik Chava. (2022). Harnessing Artificial Intelligence and Big Data for Transformative Healthcare Delivery. International Journal on Recent and Innovation Trends in Computing and Communication, 10(12), 502–520. Retrieved from https://ijritcc.org/index.php/ijritcc/article/view/11583

[38] Challa, K. (2022). The Future of Cashless Economies Through Big Data Analytics in Payment Systems. International Journal of Scientific Research and Modern Technology, 60–70. https://doi.org/10.38124/ijsrmt.v1i12.467

[39] Pamisetty, V., Pandiri, L., Annapareddy, V. N., & Sriram, H. K. (2022). Leveraging AI, Machine Learning, And Big Data For Enhancing Tax Compliance, Fraud Detection, And Predictive Analytics In Government Financial Management. Machine Learning, And Big Data For Enhancing Tax Compliance, Fraud Detection, And Predictive Analytics In Government Financial Management (June 15, 2022).

[40] Innovations in Spinal Muscular Atrophy: From Gene Therapy to Disease-Modifying Treatments. (2021). International Journal of Engineering and Computer Science, 10(12), 25531-25551. https://doi.org/10.18535/ijecs.v10i12.4659

[41] Kaulwar, P. K. (2022). Data-Engineered Intelligence: An AI-Driven Framework for Scalable and Compliant Tax Consulting Ecosystems. Kurdish Studies, 10 (2), 774–788.

[42] Operationalizing Intelligence: A Unified Approach to MLOps and Scalable AI Workflows in Hybrid Cloud Environments. (2022). International Journal of Engineering and Computer Science, 11(12), 25691-25710. https://doi.org/10.18535/ijecs.v11i12.4743

[43] Nandan, B. P., & Chitta, S. (2022). Advanced Optical Proximity Correction (OPC) Techniques in Computational Lithography: Addressing the Challenges of Pattern Fidelity and Edge Placement Error. Global Journal of Medical Case Reports, 2(1), 58-75.

[44] Raviteja Meda. (2021). Machine Learning-Based Color Recommendation Engines for Enhanced Customer Personalization. Journal of International Crisis and Risk Communication Research , 124–140. Retrieved from https://jicrcr.com/index.php/jicrcr/article/view/3018

[45] Rao Suura, S. (2021). Personalized Health Care Decisions Powered By Big Data And Generative Artificial Intelligence In Genomic Diagnostics. Journal of Survey in Fisheries Sciences. https://doi.org/10.53555/sfs.v7i3.3558

[46] Implementing Infrastructure-as-Code for Telecom Networks: Challenges and Best Practices for Scalable Service Orchestration. (2021). International Journal of Engineering and Computer Science, 10(12), 25631-25650. https://doi.org/10.18535/ijecs.v10i12.4671

[47] Vamsee Pamisetty, Lahari Pandiri, Sneha Singireddy, Venkata Narasareddy Annapareddy, Harish Kumar Sriram. (2022). Leveraging AI, Machine Learning, And Big Data For Enhancing Tax Compliance, Fraud Detection, And Predictive Analytics In Government Financial Management. Migration Letters, 19(S5), 1770–1784. Retrieved from https://migrationletters.com/index.php/ml/article/view/11808

[48] Someshwar Mashetty. (2020). Affordable Housing Through Smart Mortgage Financing: Technology, Analytics, And Innovation. International Journal on Recent and Innovation Trends in Computing and Communication, 8(12), 99–110. Retrieved from https://ijritcc.org/index.php/ijritcc/article/view/11581

[49] Srinivasa Rao Challa,. (2022). Cloud-Powered Financial Intelligence: Integrating AI and Big Data for Smarter Wealth Management Solutions. Mathematical Statistician and Engineering Applications, 71(4), 16842–16862. Retrieved from https://philstat.org/index.php/MSEA/article/view/2977

[50] Paleti, S. (2022). Fusion Bank: Integrating AI-Driven Financial Innovations with Risk-Aware Data Engineering in Modern Banking. Mathematical Statistician and Engineering Applications, 71(4), 16785-16800.

[51] Pamisetty, V. (2022). Transforming Fiscal Impact Analysis with AI, Big Data, and Cloud Computing: A Framework for Modern Public Sector Finance. Big Data, and Cloud Computing: A Framework for Modern Public Sector Finance (November 30, 2022).

[52] Kommaragiri, V. B., Gadi, A. L., Kannan, S., & Preethish Nanan, B. (2021). Advanced Computational Technologies in Vehicle Production, Digital Connectivity, and Sustainable Transportation: Innovations in Intelligent Systems, Eco-Friendly Manufacturing, and Financial Optimization.

[53] Annapareddy, V. N. (2022). Integrating AI, Machine Learning, and Cloud Computing to Drive Innovation in Renewable Energy Systems and Education Technology Solutions. Available at SSRN 5240116.

[54] Transforming Renewable Energy and Educational Technologies Through AI, Machine Learning, Big Data Analytics, and Cloud-Based IT Integrations. (2021). International Journal of Engineering and Computer Science, 10(12), 25572-25585. https://doi.org/10.18535/ijecs.v10i12.4665

[55] Venkata Bhardwaj Komaragiri. (2021). Machine Learning Models for Predictive Maintenance and Performance Optimization in Telecom Infrastructure. Journal of International Crisis and Risk Communication Research , 141–167. Retrieved from https://jicrcr.com/index.php/jicrcr/article/view/3019

[56] Paleti, S. (2021). Cognitive Core Banking: A Data-Engineered, AI-Infused Architecture for Proactive

Risk Compliance Management. AI-Infused Architecture for Proactive Risk Compliance Management (December 21, 2021).

[57] Harish Kumar Sriram. (2022). AI-Driven Optimization of Intelligent Supply Chains and Payment Systems: Enhancing Security, Tax Compliance, and Audit Efficiency in Financial Operations. Mathematical Statistician and Engineering Applications, 71(4), 16729–16748. Retrieved from https://philstat.org/index.php/MSEA/article/view/2966

[58] Chava, K., Chakilam, C., Suura, S. R., & Recharla, M. (2021). Advancing Healthcare Innovation in 2021: Integrating AI, Digital Health Technologies, and Precision Medicine for Improved Patient Outcomes. Global Journal of Medical Case Reports, 1(1), 29-41.

[59] Data Engineering Architectures for Real-Time Quality Monitoring in Paint Production Lines. (2020). International Journal of Engineering and Computer Science, 9(12), 25289-25303. https://doi.org/10.18535/ijecs.v9i12.4587

[60] Pallav Kumar Kaulwar. (2021). From Code to Counsel: Deep Learning and Data Engineering Synergy for Intelligent Tax Strategy Generation. Journal of International Crisis and Risk Communication Research , 1–20. Retrieved from https://jicrcr.com/index.php/jicrcr/article/view/2967

[61] Pandiri, L., & Chitta, S. (2022). Leveraging AI and Big Data for Real-Time Risk Profiling and Claims Processing: A Case Study on Usage-Based Auto Insurance. Kurdish Studies. https://doi.org/10.53555/ks.v10i2.3760

[62] Kummari, D. N. (2022). AI-Driven Predictive Maintenance for Industrial Robots in Automotive Manufacturing: A Case Study. International Journal of Scientific Research and Modern Technology, 107–119. https://doi.org/10.38124/ijsrmt.v1i12.489

[63] Gadi, A. L. (2022). Cloud-Native Data Governance for Next-Generation Automotive Manufacturing: Securing, Managing, and Optimizing Big Data in AI-Driven Production Systems. Kurdish Studies. https://doi.org/10.53555/ks.v10i2.3758

[64] Dodda, A. (2022). Secure and Ethical Deployment of AI in Digital Payments: A Framework for the Future of Fintech. Kurdish Studies. https://doi.org/10.53555/ks.v10i2.3834

[65] Gadi, A. L. (2021). The Future of Automotive Mobility: Integrating Cloud-Based Connected Services for Sustainable and Autonomous Transportation. International Journal on Recent and Innovation Trends in Computing and Communication, 9(12), 179-187.

[66] Dodda, A. (2022). Strategic Financial Intelligence: Using Machine Learning to Inform Partnership Driven Growth in Global Payment Networks. International Journal of Scientific Research and Modern Technology, 1(12), 10-25.

[67] Just-in-Time Inventory Management Using Reinforcement Learning in Automotive Supply Chains. (2021). International Journal of Engineering and Computer Science, 10(12), 25586-25605. https://doi.org/10.18535/ijecs.v10i12.4666

[68] Srinivasa Rao Challa. (2021). From Data to Decisions: Leveraging Machine Learning and Cloud Computing in Modern Wealth Management. Journal of International Crisis and Risk Communication Research , 102–123. Retrieved from https://jicrcr.com/index.php/jicrcr/article/view/3017

[69] Kommaragiri, V. B. (2021). Enhancing Telecom Security Through Big Data Analytics and Cloud-Based Threat Intelligence. Available at SSRN 5240140.