The Cloud, AI and Machine Learning  
CST8912 Solutions Architecture, 1/11/2024



Group 2-CST8912  
Section: 13  
Caleb Watson-Danis- 041041241

Catherine Daigle- 41175118  
Farhana Mahmud- 041008887  
Elias Ngugi Kariuki- 41114696  
Yue Gao-040895157

Professor: Madaan, Ragini

Table of Contents

[Abstract: 3](#_Toc183962679)

[Introduction: 3](#_Toc183962680)

[Literature Review 3](#_Toc183962681)

[Use Cases: 4](#_Toc183962682)

[Interviews and Surveys 6](#_Toc183962683)

[Technology Analysis 10](#_Toc183962684)

[Types of Neural Networks: 11](#_Toc183962685)

[Speech & Text services: 12](#_Toc183962686)

[Diagnostic services 12](#_Toc183962687)

[Amazon Lookout for Metrics 12](#_Toc183962688)

[Image services 13](#_Toc183962689)

[Platform services 13](#_Toc183962690)

[Azure Machine Learning 14](#_Toc183962691)

[Azure Artificial Intelligence 15](#_Toc183962692)

[Using AI Tools with Machine Learning 16](#_Toc183962693)

[Market Analysis 16](#_Toc183962694)

[Growth Forecasts 16](#_Toc183962695)

[Competitive Landscape 16](#_Toc183962696)

[Key Providers: 17](#_Toc183962697)

[Emerging Opportunities 18](#_Toc183962698)

[Netflix Case Study 18](#_Toc183962699)

[Challenges with AI/ML in the Cloud 18](#_Toc183962700)

[Challenges 19](#_Toc183962701)

[How Netflix Overcame These Challenges: 20](#_Toc183962702)

[Data-Driven Strategies for Streaming Netflix's 20](#_Toc183962703)

[Leveraging AI for Streaming Transformation 20](#_Toc183962704)

[Deep Learning for Recommender Systems 20](#_Toc183962705)

[Machine Learning in Financial Forecasting 20](#_Toc183962706)

[Ethical Challenges 20](#_Toc183962707)

[Data Privacy and Security 20](#_Toc183962708)

[Bias and Fairness 21](#_Toc183962709)

[Accountability and Explainability 21](#_Toc183962710)

[Environmental Impact 21](#_Toc183962711)

[Design Challenges and Strategies 21](#_Toc183962712)

[Scalability and Flexibility 21](#_Toc183962713)

[Model Deployment and Maintenance 22](#_Toc183962714)

[Latency and Real-Time Processing 22](#_Toc183962715)

[Interoperability and Integration 22](#_Toc183962716)

[Considerations for Implementation 22](#_Toc183962717)

[Data Governance and Compliance 22](#_Toc183962718)

[Cost Management 22](#_Toc183962719)

[Ethics Review Boards and Audits 23](#_Toc183962720)

[Human Oversight and Safety 23](#_Toc183962721)

[Conclusion 23](#_Toc183962722)

[Group Meeting Log 23](#_Toc183962723)

[References 24](#_Toc183962724)

# Abstract:

In this research, we discover how AI and Machine learning affect businesses using the cloud. We will review the challenges, Explore AI and ML (Machine Learning), look at business surveys and interviews, and observe the tools used in how AI/ML affected the business market.

# Introduction:

AI and Machine learning help enterprises scale, automate and grow with the Cloud. There are challenges and problems when managing a large enterprise by looking at Schneider Electric, Toyota, and BBC, we can see these companies apply cloud computing and machine learning to overcome these challenges.

Numerous Tools in AI and Machine Learning are used in the cloud, and we will take an in-depth look into each tool and how they operate. The tools presented are Platform services, Imaging services, Speech & Text Services and Diagnostic services. Cloud and machine learning have already affected many businesses, and surveys conducted on machine learning reveal business production has increased significantly after adopting machine learning to their cloud.

Finally, machine learning and AI technology have recently been integrated with the cloud. The future of machine learning, AI and the cloud is promising as storage options and monitoring services are growing, allowing cloud users to access big data anytime for deep learning.

# Literature Review

Artificial intelligence in cloud computing is developing quickly in different industries as more companies embrace the cloud platform. AI cloud computing is focused on building these intelligent applications, helping enterprises leverage big data, providing algorithms for advanced app functionality, and predicting future growth. (Soni & Kumar, 2023)

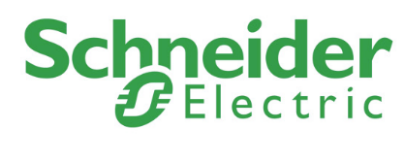
The synergy between AI, Machine Learning, and cloud services empowers IoT ecosystems to evolve beyond basic data collection, fostering intelligent automation and predictive analytics (Chen & Kumar, 2023). By integrating the cloud platform with artificial intelligence, companies can leverage the scalability of computational resources for complex Artificial intelligence algorithms.

Azure, AWS, and Google Cloud platforms are key providers, tailoring AI-driven applications by providing an environment with dedicated tools and services. The evolution of artificial intelligence in the cloud is very promising and has brought about a new era of innovation, reshaping operations in the digital world.

Use Cases:

  
[Microsoft Customer Story-Toyota Motor North America turns employee ideas into apps with Microsoft Power Platform](https://customers.microsoft.com/en-us/story/763052-toyota-motor-north-america-automotive-power-apps)

|  |  |  |
| --- | --- | --- |
| Challenges | Azure Services Used | Results |
| TMNA struggled to foster employee innovation and needed numerous applications for various functions, such as quality control and COVID-19 screenings, but lacked the technical expertise to develop them efficiently. | TMNA adopted Microsoft Azure's Power Platform, including Power Apps and Power Automate, allowing employees to create applications easily with no-code and low-code solutions. | The implementation led to the development of over 400 applications, significantly enhancing productivity and efficiency while fostering a collaborative culture of innovation supported by a Center of Excellence for governance and training. |

  
[Microsoft Customer Story-Schneider Electric fast-tracks innovation with Azure OpenAI Service](https://customers.microsoft.com/en-us/story/1745242950134216820-schneider-electric-azure-machine-learning-discrete-manufacturing-en-france?culture=en-us&country=us)Schneider Electric

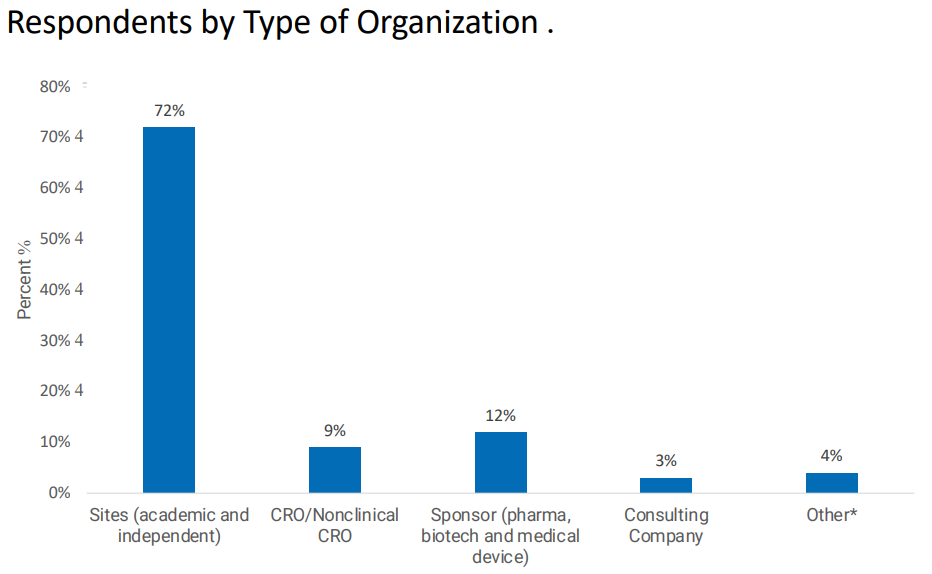
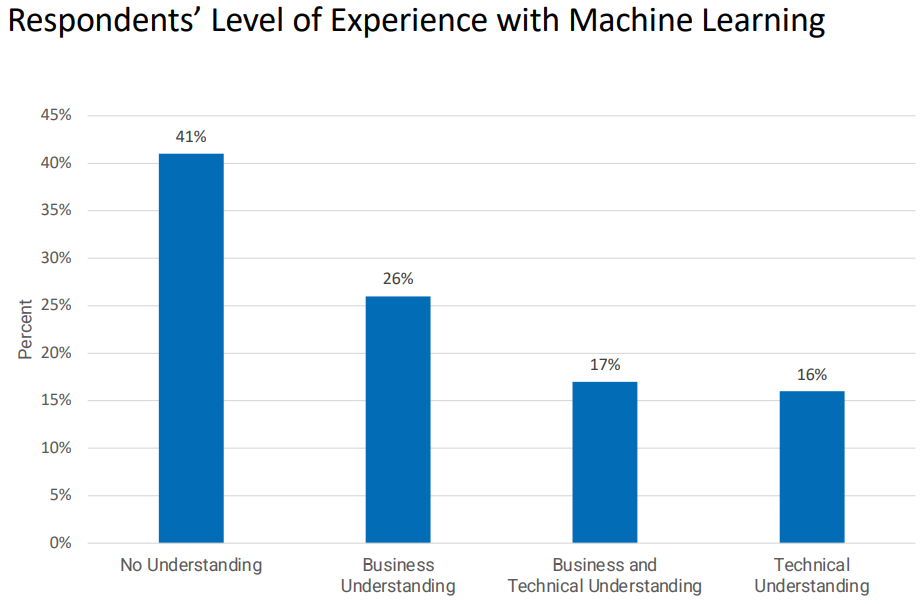
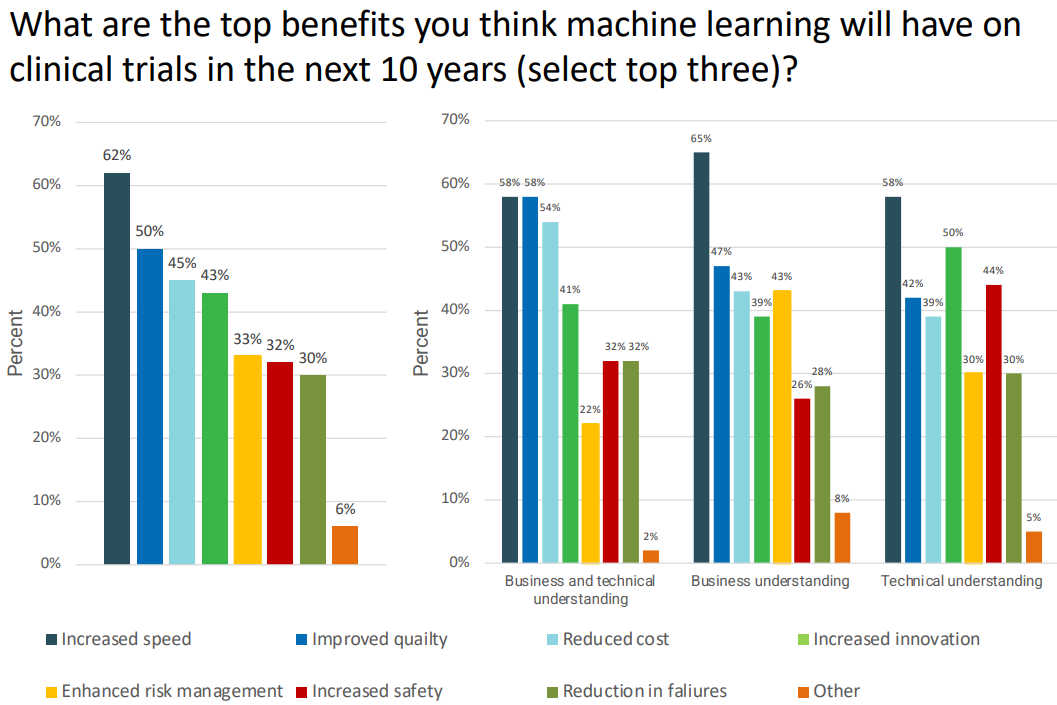
|  |  |  |
| --- | --- | --- |
| Challenges | Azure Services Used | Results |
| Schneider Electric faced the challenge of enhancing sustainability and efficiency while reducing carbon emissions.  The complexity of managing energy resources and optimizing operations for global customers required advanced technology to process large amounts of data and automate decision-making. | To address these challenges, Schneider Electric partnered with Microsoft and integrated Azure OpenAI Service, Azure Machine Learning, and other tools from Microsoft Cloud for Manufacturing.  These services helped automate tasks, optimize energy use, and improve internal efficiency. Solutions like EcoStruxure, which leverages IoT and AI, allowing Schneider to gather and analyze data for informed decision-making. | Azure services led to increased productivity, faster innovation, and enhanced sustainability efforts at Schneider. Their AI-powered solutions, such as EcoStruxure Resource Advisor, enabled customers to manage energy efficiently across global portfolios, reducing emissions and driving smarter energy usage.  These results allowed Schneider to accelerate its growth and have a positive impact on the environment. |

  
[Microsoft Customer Story-BBC innovates how it delivers trusted news and entertainment with Azure AI](https://customers.microsoft.com/en-us/story/754836-bbc-media-entertainment-azure)

|  |  |  |
| --- | --- | --- |
| Challenges | Azure Services Used | Results |
| As one of the world's leading broadcasters, the BBC faced the challenge of creating a branded voice assistant that maintained full control over customer data and relationships.  The BBC needed to provide a natural, inclusive experience for diverse audiences with various accents and dialects, making voice recognition a significant challenge.  Additionally, developing a scalable AI platform for global audiences required resources and expertise beyond what the BBC could handle alone. | To meet these challenges, the BBC utilized Microsoft Azure Cognitive Services, Azure Bot Service, and Custom Neural Voice. These services provided the flexibility and control necessary for building a natural-sounding, customized voice assistant.  Azure’s Language Understanding and Direct Line Speech further helped the BBC enhance speech recognition and integrate their systems seamlessly, ensuring the voice assistant could access a wide range of content, from podcasts to news updates. | With the help of Azure’s tools, the BBC successfully launched its Beeb voice assistant, delivering a personalized experience while maintaining privacy for users. The assistant allows audiences to discover content through voice commands, improving user engagement.  The BBC is now planning to expand the assistant’s capabilities to global markets, ensuring it can interact with users across different regions and languages. This project also emphasized the BBC's commitment to responsible AI and protecting customer data. |

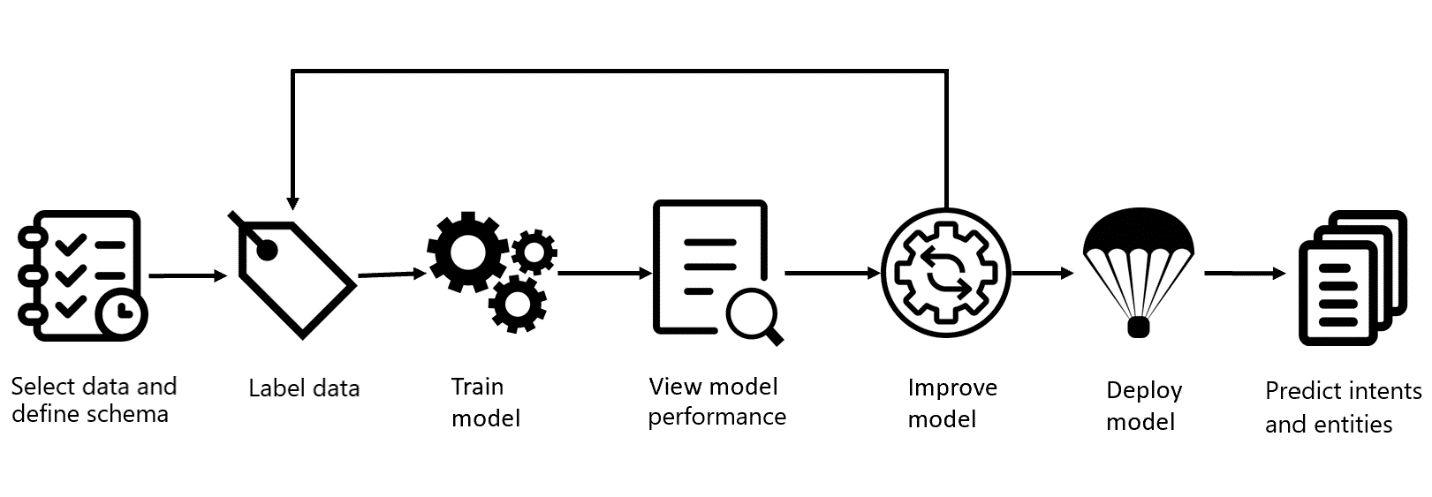
# Interviews and Surveys

The AI and Machine Learning survey report conducted by WCG CenterWatch in 2020 contains compelling information about the use of AI and machine learning within the medical sector, showing the interest level of the industry in it.  
  
The following figure shows the demographic of users that responded to the survey:

To fully understand the demographic, this is a sample size of 311 respondents: 224 from sites, 28 from CRO/Nonclinical CRO, 36 from sponsors, 11 from consulting companies and 12 from other sources.  
  
The figure below shows the level of understanding of AI and machine learning from the respondents in the survey. It is noted in the census that nearly half of the respondents (41%) have no experience with AI and machine learning. We can see that some technology sectors have little need for knowledge of machine learning, given that most respondents have worked in this industry without any knowledge or just with a business understanding of AI & ML.  
  
Now that we have the demographic of respondents and their knowledge of ML, the following are pertinent questions and their results.  
  
One intriguing question is the benefits that the respondents thought that ML might have on clinical trials:  
There are some key observations to be made about this question’s results. We can notice that across the board, regardless of understanding, the respondents agree that increased speed is a benefit of ML. However, there is a massive difference of the Increased Innovation benefit between those with technical understanding versus business understanding. 50% of those with technical understanding have Increased Innovation in their top 3 benefits, whereas it drops to 39% for those with only business understanding. This can likely be attributed to those with technical understanding having a deeper knowledge of machine learning’s capabilities and understanding how it could evolve and help in their industry.   
  
Another key change that can be noted is Increased Safety. 44% of those with technical understanding chose it as a benefit, whereas only 26% of those with business understanding chose it. This makes sense because we want to believe that safety is within our legacy systems, and we grow accustomed to them. However, machine learning can greatly increase safety using ML algorithms to pinpoint malicious patterns or encrypted traffic (<https://www.cisco.com/c/en/us/products/security/machine-learning-security.html#~how-ml-helps-security>). It makes sense that those with technical understanding (44%) are more eager to believe in the increased security than those with business understanding (26%).

# Technology Analysis

The cloud offers a variety of tools and technologies to support the various aspects of machine learning and AI. Each tool may only differ slightly between cloud providers. The key providers of such tools are AWS, Azure, and GCP. Speech & Text services, Diagnostic services, Image services and Platform services are common tools offered by cloud providers. These tools can be adapted to machine-learning algorithms to speed production, conserve costs, and automate applications.

This diagram shows the basic workflow for deploying common machine-learning tools such as Speech & Text, Diagnostic and Image services. Each step is as follows: (*What is conversational language understanding?* 2024)

* **Define the schema:** Define the information data needed for your model to learn and define what you want your machine learning model to predict. For example; Netflix recommends movies from user data.
* **Label datasets:** Labelling is a key aspect of Machine learning; it defines what the AI should look for in a dataset and the meaning behind the data. For example, a pet store may want to label images that contain dogs as “dogs.”
* **Training the model:** Training the A.I. algorithm allows the algorithm to learn “correct” patterns using the previous step’s labels.
* **View performance:** Assess the algorithm’s performance and confidence when with new, unlabeled data. If the model is not confident in its prediction, you may need to return to re-labelling the dataset. Confidence defines how accurate a model’s prediction is.
* **Improving the model:** Optimize the model’s performance and confidence levels for better predictions of the meaning behind data.
* **Deploying the model:** Prepare the model and its dependencies to be usable within the cloud. Here, you build the model and deploy it to production as “beta” for testing.
* **Predict intent:** Use the algorithm to predict relevant information from user input. And test the product. The model should be operating efficiently for users at this step.   
  (*What is conversational language understanding?* 2024)

### Types of Neural Networks:

Before touching on the various services offered by cloud providers, one must understand how these services are made and what type of neural network functions within them. This is important especially when using a custom NN workshop such as Azure Machine Learning to develop customized NN models. Below is an image representing the different types of NN:

A diagram of a network

Description automatically generated

(Sahota, 2022)

* **Feed Forward NN**: These NN are the most basic NN type and may be useful for small applications that require a lightweight NN.
* **Convolutional NN**: These NN are best use for images for their way of processing data in grid structures. Images are formatted as a pixelized grid which is why these NN are best.
* **Recurrent NN:** Recurrent NN are best for small translation applications as they designed in a way to remember previous text structure. However, they do not have the same complexity and accuracy as auto encoder NN to specific text remembrance.
* **Generative Adversarial NN:** Involves multiple NN working against one another to determine a prediction. They are best for processing large amounts of data and creating images like DALL-E.
* **Auto-Encoder NN:** This NN specializes in “remembering” specific inputs and applying it in training over long sequences of data. Auto-Encoders are best for speech recognition software such as Chat GPT. (Sahota, 2022)

## Speech & Text services:

Speech services allow cloud users to use or build AI language models to analyze and predict customer documents and text prompts. AI Speech services can include Speech-to-text, Text-to-speech, Chatbots, Translation and Document Analysis.

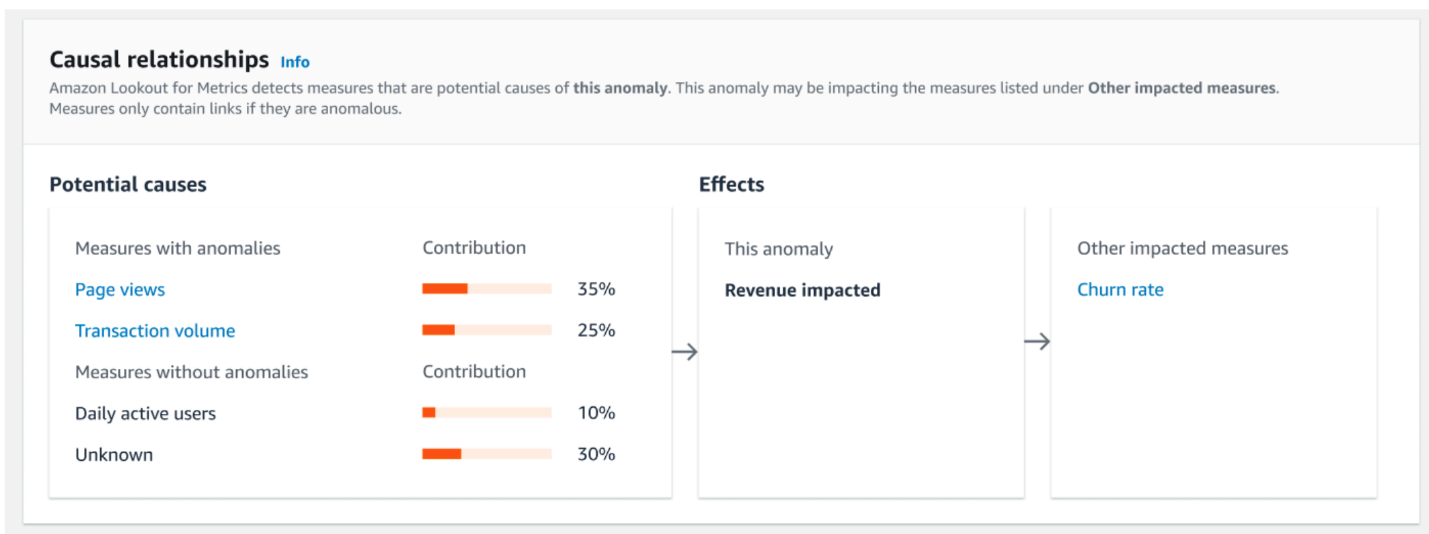
## Diagnostic services

The cloud can also provide diagnostic services for web applications such as Azure AI Metrics Advisor or Amazon Lookout. These services analyze data, detect anomalies, and diagnose web workspaces.

### Amazon Lookout for Metrics

Lookout for Metrics is Amazon’s version of diagnostic services. Within Lookout, you create and configure resources to identify changes in data within your business applications, called anomalies. You can set these resources to analyze your data within a timed schedule or specify a location. When a resource finds an anomaly, it can notify you and output a graph of what caused the anomaly and what it affects.

(*Amazon Lookout for Metrics - Developer Guide* 2024)

(*Amazon Lookout for Metrics - Developer Guide* 2024)  
  
Microsoft Azure may also output similar graphs that show the relationship between what caused the anomalies and what data got impacted. Azure AI Metrics Advisor and Amazon Lookout use AI to diagnose Anomalies and determine what caused them with only slight differences. For example, the resources to detect anomalies have different names: Amazon calls them” detectors,” and Microsoft calls them” Measures.” (*How-to: Onboard your metric data to Metrics Advisor* 2024)

## Image services

The cloud provides AI imaging tools to deploy into larger-scale image recognition models. This tool uses computer vision to extract information from faces, texts, images and video for use in any application that requires image recognition. Azure calls this service Azure AI Vision, while Amazon’s is Image Rekognition. (*What is Azure AI Vision?* 2024)

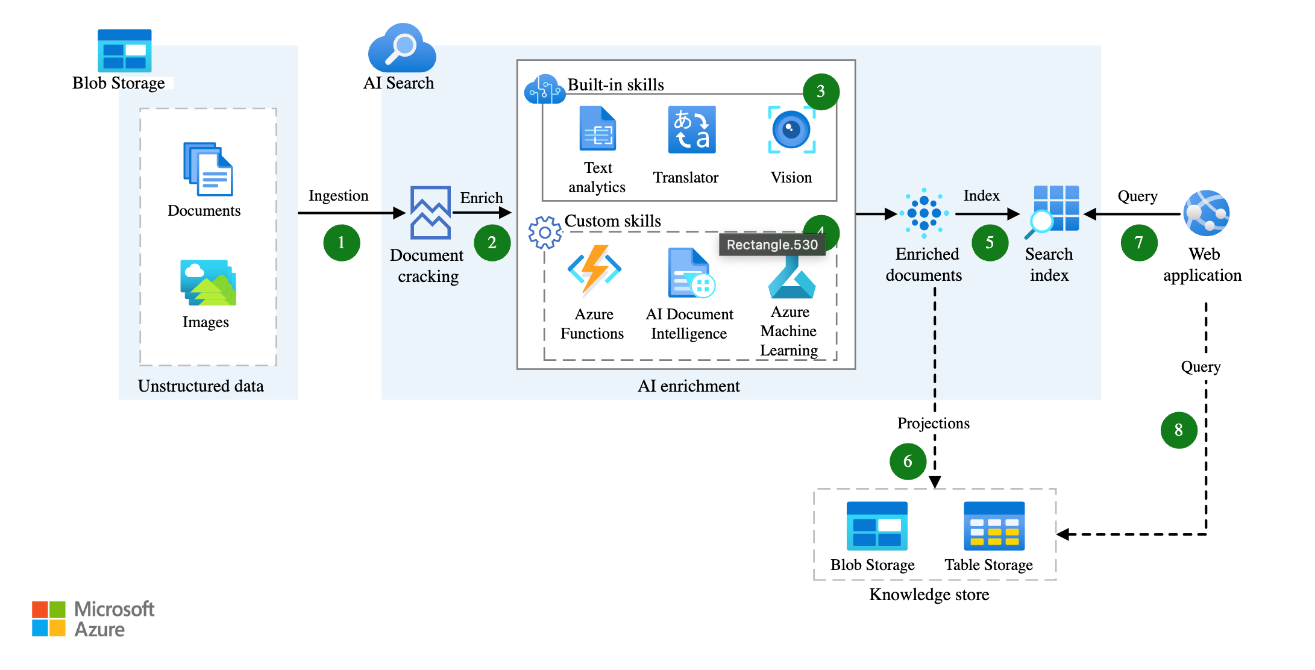
## Platform services

Each cloud provider may integrate outside machine learning platform services into the cloud. For example, Amazon and Azure use ML frameworks like Py Torch, TensorFlow or scikit-learn. Platform services like Azure Machine Learning provide fully integrated workspaces to develop and manage machine learning projects.

### Azure Machine Learning

Azure Machine learning integrates everything you need for teams of developers, engineers and data scientists to build fully scalable ML models for production and automation. Azure ML implements Jupyter Notebook for writing ML code and uses Py Torch or TensorFlow for building training models. For security, Azure ML uses Azure virtual networks and Azure Key Vault to store sensitive information.

Below is a Cloud-based Machine Learning Image architecture using AI and Machine Learning Technology:



(Santos, *Use AI enrichment with image and text processing* 2023)

In this architecture diagram, the cloud combines AI and Machine learning to predict the meaning behind images and text. The AI tools use Speech & Text and image services such as Azure Translator, Text analytics and Azure AI Vision. Here, Azure Machine Learning supports specific use cases and scenarios. The machine learning architecture is split into Built-in skills, pre-built AI tools and models and Custom skills for specific custom ML models. (*What is Azure Machine Learning?* 2024)

## Azure Artificial Intelligence

A diagram of a software process

Description automatically generated

(Microsoft. (2023, August 21). Extract and analyze call center data.)

This architecture extracts call data, analyzes it, and uses PowerBI to visualize the data for the optimization of call centers. This architecture does not focus on machine learning rather, it focuses on how AI services can work together to create call-center analytics. Machine learning is mostly used for predictive analytics and can be complex to deploy. Using AI services can simplify the architecture and reduce costs, as you can deploy or shut down the AI services according to web traffic.

* First, the call is extracted from the customer using the Telephony server.
* The call data is then stored within an Azure Blob storage to analyze.
* Azure then converts the call into a text transcript using Azure speech to text.
* The transcript data is then stored in Azure blob storage.
* Azure AI language detects and removes personal information from the transcript.
* Azure OpenAI is used to analyze the transcript and shorten it. The data is then stored once again within Azure blob.
* Power Bi provides visual insights into the data in real-time displaying the information with graphs, pie charts and bar graphs.

## Using AI Tools with Machine Learning

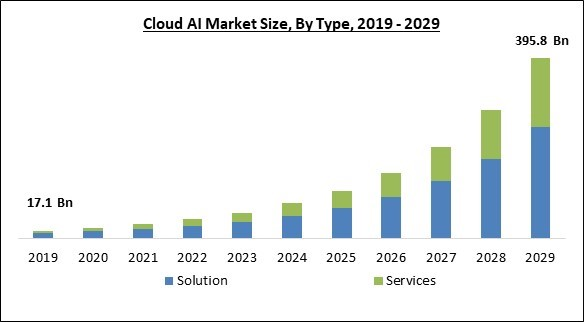
With pre-packaged AI tools such as Image services, diagnostic services and speech/text services, a business can combine these services with a Platform service such as Azure Machine Learning to create and deploy an expansive machine learning model that can account for specific use cases without the additional work to produce these tools.

# Market Analysis

When combining cloud computing with artificial intelligence (AI) and machine learning (ML) is an important thing for the market because it makes competition greater and creates new opportunities for new businesses. This section is an in-depth market analysis, which include the predicted growth forecasts, competitive landscape and emerging opportunities.

## Growth Forecasts

When we look at the growth of AI, it’s becoming an exponentially growing sector that is pushing the bounds of cloud development. It’s noted that the Global Cloud AI Market size is expected to reach $395.8 billion by 2029, rising at a market growth of 38.4% CAGR during the forecast period. (Marqual IT Solutions Pvt. Ltd,2023) This is a staggering increase in market size, and shows how present the global interest is there for AI. The figure below shows these statistics:



## Competitive Landscape

The competitive landscape for AI and ML services is always challenging each other frequently, innovation is a primary factor that drives the competitive landscape of the cloud AI market to help expand the reach of these services markets. Leading cloud providers such as Amazon’s AWS, Microsoft’s Azure, and Google’s GCP lead the market with their powerful AI capabilities. These companies are trying to enhance AI algorithms, improve data security, and optimize the infrastructure that makes up cloud AI.

### Key Providers:

This section introduces some of the major cloud vendors and their features and offerings.

|  |  |  |
| --- | --- | --- |
| Cloud Provider | AI/ML Offering | Key Strengths |
| Amazon Web Services (AWS) | Amazon SageMaker | - Powerful platform for building, training, and deploying large-scale ML models; enhances ML productivity.  - Extensive integration with other AWS services.  - Strong support for developers with tools and documentation. |
| Google Cloud | Vertex AI | - Unified platform for training and deploying ML models; allows customization of large language models (LLMs).  - Strong data analytics capabilities.  - Seamless integration with other Google services. |
| IBM | Watsonx | - Enterprise-focused AI platform for advanced analytics, ML, and scalable AI workflows; strong emphasis on data.  - Strong support for industries like healthcare and finance. |
| Microsoft Azure AI | Azure AI | - Comprehensive tools for building, training, and deploying ML models; integrates seamlessly with intelligent apps.  - Robust enterprise support and security features.  - User-friendly interfaces and tools. |
| Salesforce | Einstein AI: Integrated AI for CRM | - Seamless integration with Salesforce ecosystem.  - Tailored AI solutions for sales, service, and marketing.  - User-friendly tools for non-technical users. |
| Oracle | Oracle AI Toolset | - Strong integration with Oracle databases and applications.  - Focus on enterprise-grade solutions.  - Emphasis on data security and governance. |

## Emerging Opportunities

Cloud providers and AI software developers are often collaborating in different industries, such as healthcare, finance, and manufacturing just to name a few. They help customize and create cloud solutions based on specific needs of these industries that help us every day.

|  |  |
| --- | --- |
| Industry | Use |
| Healthcare | AI-driven cloud solutions are becoming increasingly popular in telemedicine, medical imaging, and drug discovery. The healthcare sector presents a high-growth opportunity, with AI helping to improve patient outcomes and reduce costs. |
| Finance | The application of AI in fraud detection, credit scoring, and risk assessment through cloud platforms is growing rapidly. AI will enable real-time risk analysis and personalized banking, creating opportunities for cloud providers specializing in FinTech solutions. |
| Retail and E-Commerce | Cloud-based AI tools can optimize construction schedules, reduce costs, and predict maintenance needs based on 3D models. |
| Environmental Monitoring | Cloud AI can process large datasets from global monitoring systems to predict climate change patterns and their impacts on ecosystems. |

# Netflix Case Study

## Challenges with AI/ML in the Cloud

**Context:** Netflix has been a pioneer in using AI and ML to power its recommendation engine, optimize streaming quality, and predict user behavior. As the company expanded globally, they faced several difficulties in scaling their AI/ML workloads in the cloud.

## Challenges

1. **Data Integration:**
   * **Issue:** Netflix had vast amounts of data collected from different regions, devices, and user interactions. Integrating this data into their cloud infrastructure was complex due to its volume, variety, and the need for real-time processing.
   * **Impact:** Delays in data integration slowed down training machine learning models and made it harder to generate accurate recommendations.
2. **Scalability and Resource Management:**
   * **Issue:** Netflix's AI models required significant computational resources for training and real-time inference. Training deep learning models at scale is resource-intensive, and optimizing infrastructure costs while ensuring performance was challenging.
   * **Impact:** Inefficient resource allocation led to high operational costs, especially during peak periods. It became difficult to balance demand spikes and maintain real-time recommendations without overspending on infrastructure.
3. **Model Training and Deployment:**
   * **Issue:** Netflix needed to frequently update its recommendation algorithms and deploy new models. However, coordinating model training, testing, and deployment across various teams globally introduced complexity.
   * **Impact:** The lag in deploying updated models affected the quality of personalized recommendations, resulting in lower user engagement at times.
4. **Security and Compliance:**
   * **Issue:** As Netflix expanded into new markets, it had to comply with different privacy and data protection laws (e.g., GDPR in Europe). Ensuring secure data handling, while training AI models in the cloud, added to the complexity.
   * **Impact:** Compliance requirements forced Netflix to rethink its data management strategies, especially when handling user data in AI training processes.
5. **Data Drift:**
   * **Issue:** User preferences and viewing patterns changed frequently, which caused **data drift**—where the data used for training models no longer reflected current trends.
   * **Impact:** Without constant monitoring and retraining, Netflix’s recommendation algorithms would become less effective over time, leading to lower satisfaction with recommendations.

## How Netflix Overcame These Challenges:

### Data-Driven Strategies for Streaming Netflix's

* **Personalized Recommendations**: Netflix's recommendation engine makes content-based, collaborative, and deep learning-based content recommendations for each user. Because of this customisation, recommendations now account for 80% of the content that is viewed on Netflix, saving the firm about $1 billion a year (Steck et al., 2021).
* Users expressed high satisfaction with the relevance of recommendations, leading to increased viewing time and subscriber retention (Steck et al., 2021).

### Leveraging AI for Streaming Transformation

* **Adaptive Bitrate Streaming (ABR):** ABR technology dynamically modifies the quality of video broadcasts according to the user's device capabilities and internet speed using machine learning. Even with slower connections, this guarantees a buffer-free, fluid viewing experience (Veerala, n.d.).

### Deep Learning for Recommender Systems

* **Neural Collaborative Filtering:** To provide more accurate and relevant recommendations, Netflix has adopted neural collaborative filtering algorithms to capture detailed user-item interactions (Steck et al., 2021).

### Machine Learning in Financial Forecasting

* **Regression Models**: Netflix uses regression models to analyze historical stock price data and identify trends drove Netflix to have accurate stock price forecasts and recommendations (Bhati, n.d.).

# Ethical Challenges

## Data Privacy and Security

* **Data Ownership:** Since ML and AI require large datasets to be effective, deciding who owns the data, especially if it involves personal information, is crucial.
* **Data Security:** Hosting sensitive data on the cloud increases the risk of breaches. Effective encryption, strict access controls, and robust authentication are necessary to protect data.
* **User Consent and Transparency:** It's essential to inform users of how their data is used, stored, and processed. Ethical AI systems prioritize user consent and aim for transparency.

## Bias and Fairness

* **Data Bias:** AI models can unintentionally reflect or amplify biases present in the data, leading to unfair treatment of certain groups. Ensuring diverse, representative datasets is essential to mitigate this.
* **Algorithmic Fairness:** It’s important to design algorithms that prevent discrimination and ensure fair outcomes. Regular audits and testing for bias are key parts of responsible AI practices.

## Accountability and Explainability

* **Black Box Models:** Many AI models, particularly deep learning systems, lack transparency in how they make decisions, creating challenges in accountability.
* **Explainable AI (XAI):** Emphasizing interpretability and building models that can explain their decisions fosters trust and enables better error handling, especially in critical applications like healthcare.

## Environmental Impact

* **Energy Consumption:** ML and AI models, especially large-scale ones, consume substantial computational resources, contributing to carbon emissions. This can be addressed by optimizing models and choosing energy-efficient cloud providers.

## Design Challenges and Strategies

### Scalability and Flexibility

* **Elastic Infrastructure:** Cloud platforms offer the flexibility to scale up or down depending on workload, which is essential for handling varying data processing and training requirements in ML and AI applications.
* **Resource Allocation:** Optimally distributing compute resources, storage, and memory can be challenging. Automated scaling and monitoring tools help manage costs and ensure efficient resource usage.

### Model Deployment and Maintenance

* **Continuous Integration and Continuous Deployment (CI/CD):** Effective CI/CD pipelines can streamline the deployment of models, automate updates, and handle version control.
* **Monitoring and Performance Tuning:** Cloud-based AI applications need constant monitoring to detect and correct issues, such as data drift or model degradation, and maintain performance.

### Latency and Real-Time Processing

* **Low-Latency Solutions:** Some applications, like autonomous driving or real-time video analytics, require fast response times. Cloud architectures need to be designed for minimal latency, often involving edge computing to bring processing closer to the data source.

### Interoperability and Integration

* **Data and Model Portability:** Ensuring ML and AI models can be easily transferred across cloud environments is essential for flexibility and avoiding vendor lock-in.
* **API Compatibility:** Standardized APIs help make different cloud-based AI solutions interoperable and enable smoother integration across systems.

## Considerations for Implementation

### Data Governance and Compliance

* **Compliance with Regulations:** ML and AI in the cloud must comply with data protection regulations (e.g., GDPR, HIPAA) to ensure lawful handling of user data.
* **Data Provenance and Tracking:** Documenting data sources and transformations is necessary for maintaining data integrity and ensuring accountability.

### Cost Management

* **Optimizing Cloud Usage:** Cloud resources can be expensive, especially with large-scale AI models. Using tools for cost monitoring and budgeting is crucial.
* **Efficient Model Training:** Training models on large datasets can drive up costs. Techniques like transfer learning, model compression, and pruning help optimize resource usage.

### Ethics Review Boards and Audits

* **Ethical Oversight:** Regular reviews and audits help maintain ethical AI practices. These reviews can assess for fairness, inclusivity, and safety.
* **Stakeholder Inclusion:** Including diverse stakeholders in the development and deployment phases helps address ethical concerns and aligns AI applications with broader social values.

### Human Oversight and Safety

* **Human-in-the-Loop (HITL) Systems:** Ensuring human oversight in decision-making processes can help catch errors and ensure AI decisions are contextually appropriate.
* **Red Teaming and Adversarial Testing:** Testing models against potential adversarial attacks or misuse scenarios is essential to improve resilience and safety.

By carefully addressing these ethical, design, and implementation considerations, cloud-based AI solutions can be deployed more responsibly and effectively, ensuring they benefit users and society while minimizing potential harms.

Conclusion

We found the various ways AI or Machine learning can be applied to the cloud and affect businesses. AI or Machine learning can automate and provide encapsulated tools for enterprises to deploy quickly. We found from clinical trials that most enterprises do not understand machine learning but are interested in adopting it into their systems. Additionally, using AI and Machine Learning increases production speed and security significantly.

Group Meeting Log

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Meeting** | **Date** | **Caleb Watson-Danis** | **Catherine Daigle** | **Elias Ngugi Kariuki** | **Yue Gao** | **Attendnace Notes** |
| **Initial Zoom meeting** | Sept 23 | Present | Present | Present | Present | Everyone contributed, Work distribution completed |
| **In person meeting** | Sept 28 | Present | Present | Present | Present | On the track! |
| **In person meeting** | Oct 4 | Present | Present | Present | Present | Happily, contributed, everyone is ready for the presentation #1 |
| **Zoom meeting** | Oct 6 | Present | Present | Present | Present | On the right track for the research assignment1. |
| **Zoom meeting** | Nov 3 | Present | Present | Present | Present | Happily, contributed, everyone is ready for the research assignment 2. |
| **In person meeting** | Oct 17 | Present | Present | Present | Present | Happily, contributed, everyone is ready for the presentation #2 and research assignment 3. |
| **In person meeting** | Nov 8 | Present | Present | Present | Present | Happily, contributed, everyone is ready for the presentation #3 |
| **In person meeting** | Nov 22 | Present | Present | Present | Present | Happily, contributed, everyone is ready for the presentation #4 |
| **Zoom meeting** | Nov 30 | Present | Present | Present | Present | Happily, contributed, everyone is ready for the research assignment 4. |

# References

1. Amazon Web Services. (2024). *Amazon Lookout for Metrics - Developer Guide*. What is Amazon Lookout for Metrics? <https://docs.aws.amazon.com/lookoutmetrics/latest/dev/lookoutmetrics-welcome.html>
2. Bhati, Z. (n.d.). Netflix-Case-Study. GitHub. Retrieved from https://github.com/ZuhairBhati/Netflix-Case-Study
3. Cisco. (n.d.). *How machine learning enhances security*. Cisco. <https://www.cisco.com/c/en/us/products/security/machine-learning-security.html#~how-ml-helps-security>
4. Digital Socius. (2023). *How Netflix leverages big data analytics and cloud computing for global success*. Retrieved from <https://www.digitalsocius.com>
5. Lo Piano, S. (2020, June 17). *Ethical principles in machine learning and artificial intelligence: Cases from the field and possible ways forward*. Nature News. <https://www.nature.com/articles/s41599-020-0501-9>
6. Marcus Law. (2023, December 1). *How AI is changing the cloud landscape*. Technology Magazine. <https://technologymagazine.com/articles/how-ai-is-changing-the-cloud-landscape>
7. Microsoft. (2024, September 19). *What is Azure Machine Learning?*. Azure Machine Learning documentation. <https://learn.microsoft.com/pdf?url=https%3A%2F%2Flearn.microsoft.com%2Fen-us%2Fazure%2Fmachine-learning%2Ftoc.json%3Fview%3Dazureml-api-2>
8. Microsoft. (2024, August 28). *What is Azure AI Vision?*. Azure AI Vision documentation. <https://learn.microsoft.com/pdf?url=https%3A%2F%2Flearn.microsoft.com%2Fen-us%2Fazure%2Fai-services%2Fcomputer-vision%2Ftoc.json>
9. Microsoft. (2024, August 28). *How-to: Onboard your metric data to Metrics Advisor*. Azure AI Metrics Advisor documentation. <https://learn.microsoft.com/pdf?url=https%3A%2F%2Flearn.microsoft.com%2Fen-us%2Fazure%2Fai-services%2Fmetrics-advisor%2Ftoc.json>
10. Microsoft. (2024, August 28). *What is conversational language understanding?*. Azure AI Language documentation. <https://learn.microsoft.com/pdf?url=https%3A%2F%2Flearn.microsoft.com%2Fen-us%2Fazure%2Fai-services%2Flanguage-service%2Ftoc.json>
11. Microsoft. (2023, August 21). Extract and analyze call center data. Artificial intelligence (AI) architecture design. <https://learn.microsoft.com/pdf?url=https%3A%2F%2Flearn.microsoft.com%2Fen-us%2Fazure%2Farchitecture%2Fai-ml%2Ftoc.json>
12. Netflix Technology Blog. (n.d.). *Metaflow: Netflix's AI/ML infrastructure to accelerate innovation*. Retrieved from <https://netflixtechblog.com>
13. Pavan Ogeti. (2024). Benefits and Challenges of Deploying Machine Learning Models in the Cloud. *International Journal of Intelligent Systems and Applications in Engineering*, *12*(22s), 194–209.
14. Research and Markets. (2023, April). Global cloud AI market size, Share & industry trends analysis report by type, by industry, by technology (solution deep learning, machine learning, Natural Language Processing), by regional outlook and forecast, 2023-2029. Research and Markets - Market Research Reports <https://www.researchandmarkets.com/reports/5806529/global-cloud-ai-market-size-share-and-industry>
15. Santos, C. A. (2023, August 21). *Use AI enrichment with image and text processing*. Artificial intelligence (AI) architecture design. <https://learn.microsoft.com/pdf?url=https%3A%2F%2Flearn.microsoft.com%2Fen-us%2Fazure%2Farchitecture%2Fai-ml%2Ftoc.json>
16. Sahota, N. (2022, December 25). *Neural networks: A solution for complex problems in science and engineering*. Neural Networks: A Solution for Complex Problems in Science and Engineering. <https://www.neilsahota.com/neural-networks-a-solution-for-complex-problems-in-science-and-engineering/#:~:text=Neural%20networks%20rely%20heavily%20on,on%20new%20or%20unseen%20data>.
17. Steck, H., Baltrunas, L., Elahi, E., Liang, D., Raimond, Y., & Basilico, J. (2021). Deep learning for recommender systems: A Netflix case study. AI Magazine, 42(3), 7-18. <https://doi.org/10.1609/aimag.v42i3.18140>
18. Soni, D., & Kumar, N. (2023). Artificial Intelligence in cloud computing. *Artificial Intelligence in Cyber-Physical Systems*, 51–73. https://doi.org/10.1201/9781003248750-4
19. WCG CenterWatch. (2021, February). *Artificial intelligence and machine learning in life sciences: Survey report*. Oracle. <https://www.oracle.com/a/ocom/docs/industries/life-sciences/ls-ai-machine-learning-survey-report.pdf>
20. Veerala, H. K. (n.d.). Case-Study-ML-Netflix-Movie-Recommendation-System. GitHub. Retrieved from <https://github.com/veeralakrishna/Case-Study-ML-Netflix-Movie-Recommendation-System>