Scalable AI Pipelines with cloud services

Group 8

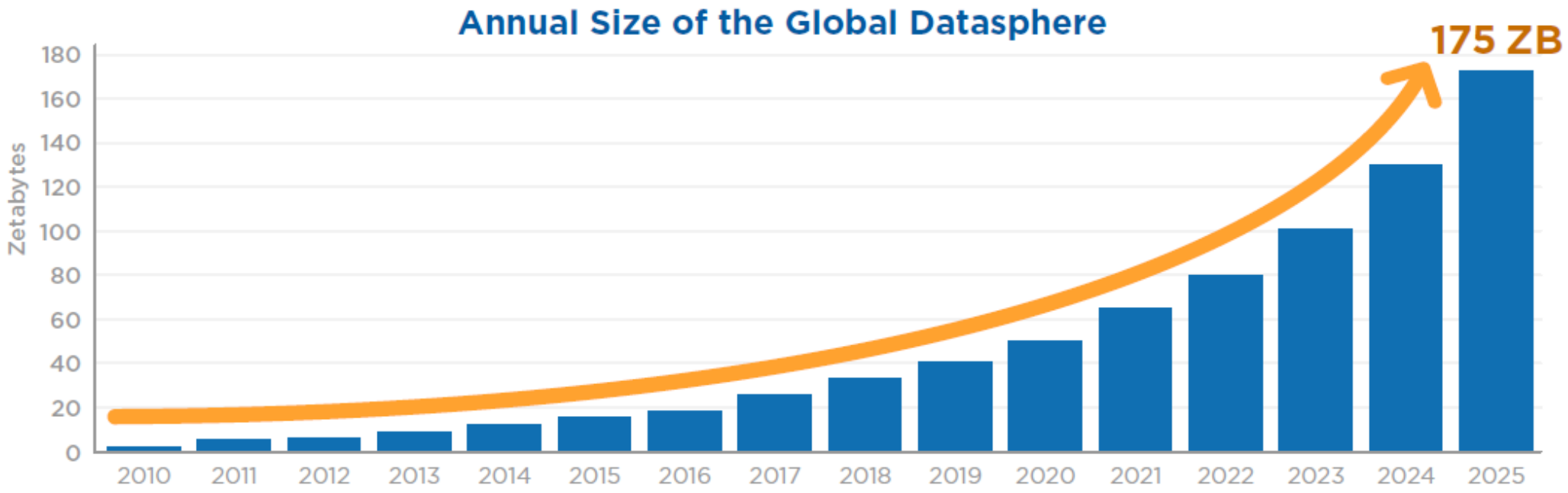
Saad Abdullah, Lameya Islam, Tirthendu Chakravorty, Abdul Wahab, Ishara Madhavi Galbokka Hewage

AGENDA

- *Why Cloud for AI/ML?*
- *Architecture for AI/ML pipelies using different cloud providers.*
- *Comparative Analysis on Cloud Providers*
- *Case Study for a real-world cloud AI pipeline.*
- *CI/CD/CT/CM in AI/ML Pipelines & Best Practices*
- *Challenges & Conclusion*

Did you know?

Figure 1 – Annual Size of the Global Datasphere



By 2025, the world is expected to generate a staggering **175 zettabytes** of data. To put that in perspective, that's 175 trillion gigabytes—enough to stack Blu-ray discs to the moon and back **23 times**!

Woodie, A. (2018, November 27). *Global DataSphere to hit 175 zettabytes by 2025, IDC says*. BigDATAwire. https://www.datanami.com/2018/11/27/global-datasphere-to-hit-175-zettabytes-by-2025-idc-says/

WHY CLOUD FOR AI/ML?

Need for Cloud in AI Pipelines?

High computational demand in AI/ML.

Traditional on-premise limitations (cost, scalability).

Reproducibility (Scaling Research)

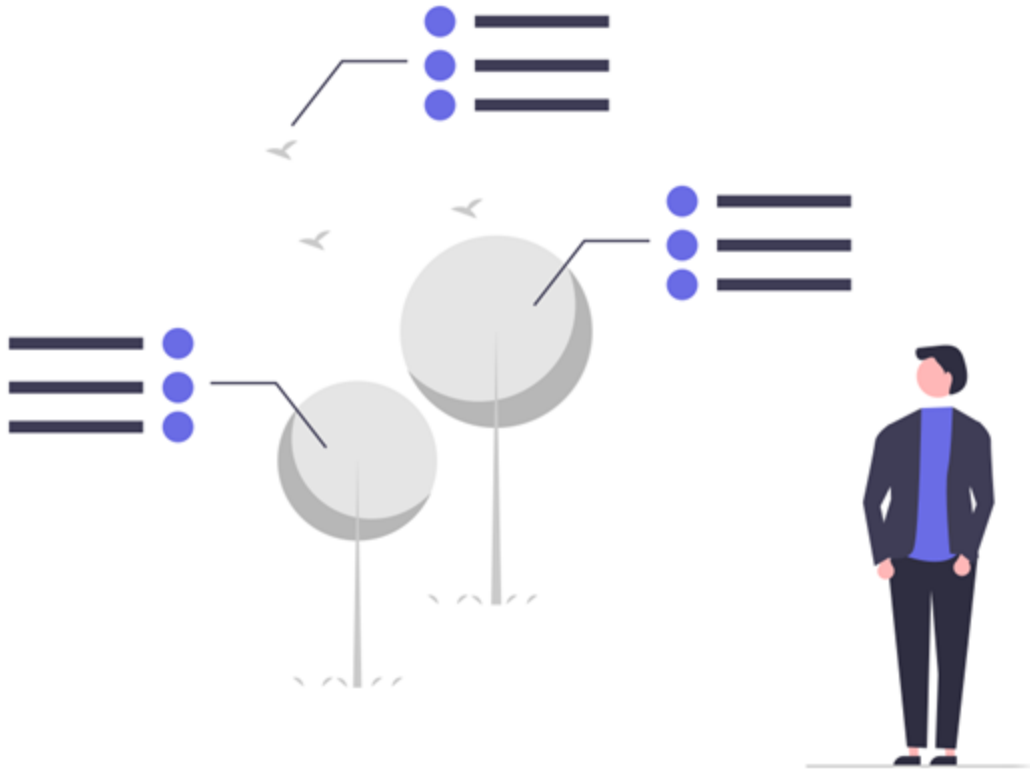
Challenges Before Cloud Adoption

Issues with Traditional AI Development?

Expensive Hardware: High costs of GPUs/TPUs, storage limitations

Inflexibility: Slow scaling, hard to accommodate increasing workloads

Data Challenges: Inefficient handling of large datasets



Cloud’s Impact on AI Pipelines

Cloud Solutions for AI/ML!

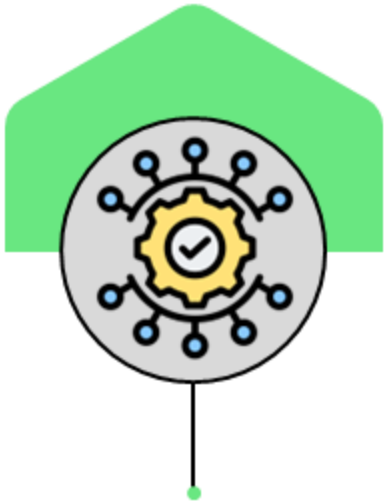
Scalability: Access to vast computational resources on demand

Collaboration: Cloud enables easier data sharing and global access

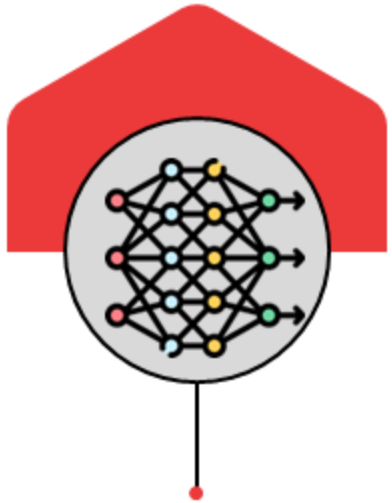
Cost-Effectiveness: Pay-as-you-go model

A Traditional ML Pipeline

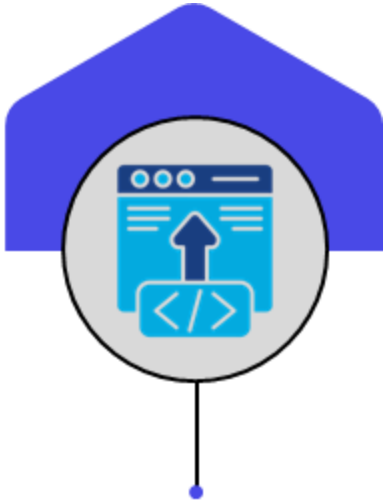
Data Storage and Management



Preprocessing and Transformation



Model Training



Model Deployment

Analysis of Cloud Services for Scalable AI/ML Pipelines

Data Storage and Management

- **Ingestion:** AWS Glue, Google Dataflow, Azure Data Factory.
- **Raw Storage**: AWS S3, Google Cloud Storage, Azure Blob Storage.

Analysis of Cloud Services for Scalable AI/ML Pipelines

Data Preprocessing and Transformation

- **ETL Pipeline:** AWS EMR, Google Dataproc, Azure HDInsight.
- **Serverless Processing:** AWS Lambda, Google Cloud Functions, Azure Functions

Analysis of Cloud Services for Scalable AI/ML Pipelines

Model Training

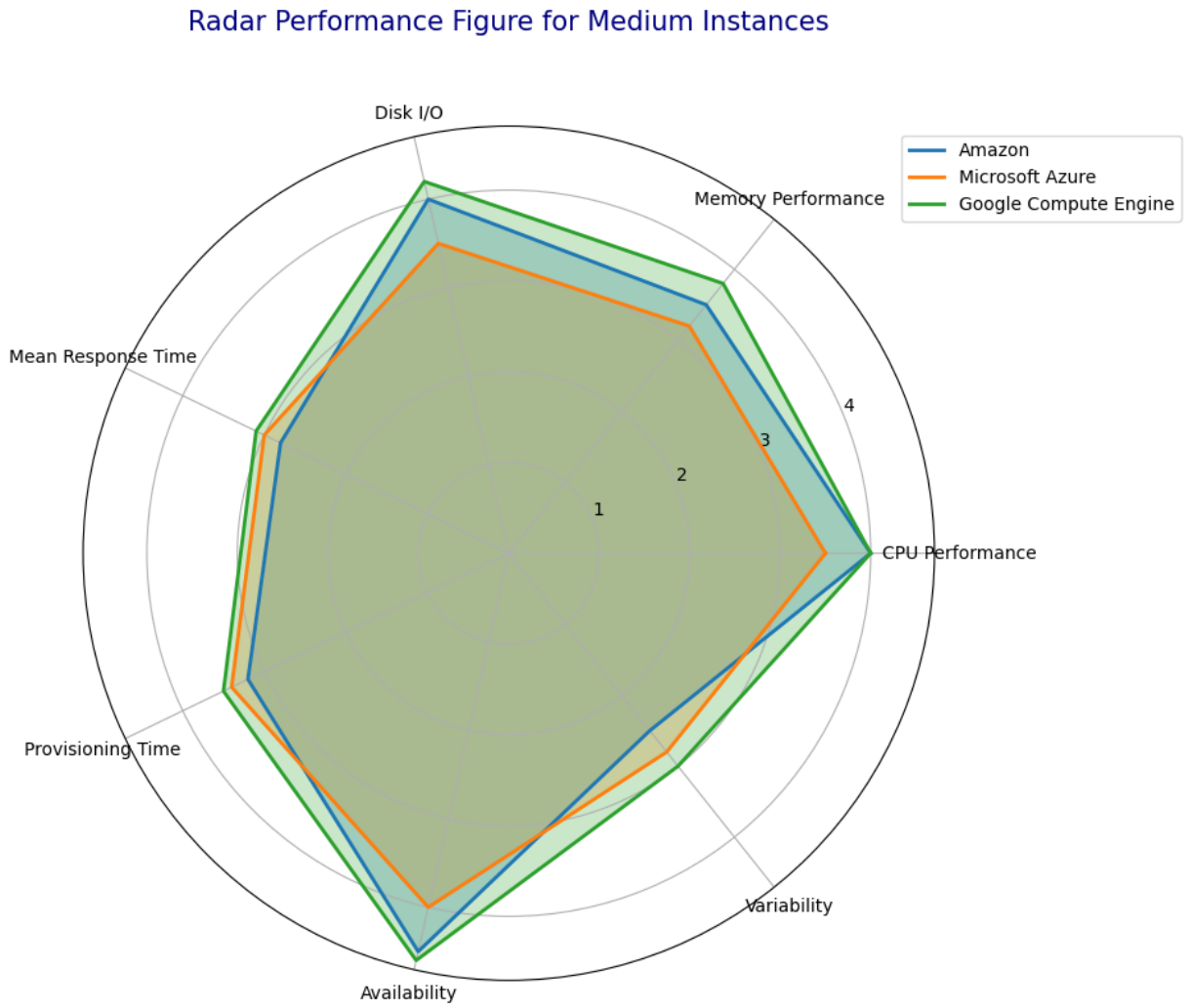
- **Distributed Training:** AWS SageMaker, Google AI Platform, Azure Machine Learning.
- **Hyperparameter Tuning:** SageMaker Hyperparameter Tuning, Google AI Platform Vizier, Azure HyperDrive

Analysis of Cloud Services for Scalable AI/ML Pipelines

Model Deployment

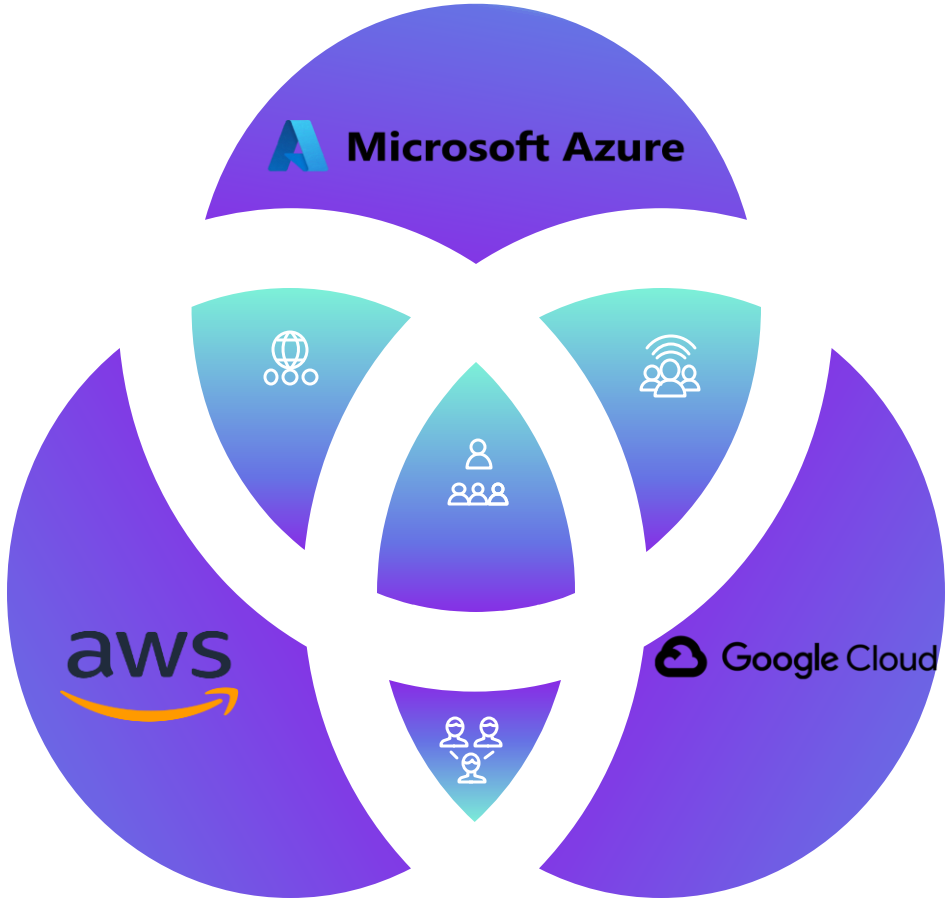
- **Real-Time Inference:** AWS SageMaker Endpoints, Google AI Platform Predictions, Azure ML Endpoints.
- **Batch Inference:** AWS Batch, Google Dataflow, Azure Batch.

Comparative Analysis



Source: [Performance and Price analysis for Cloud Service Providers]

Performance on Model Training

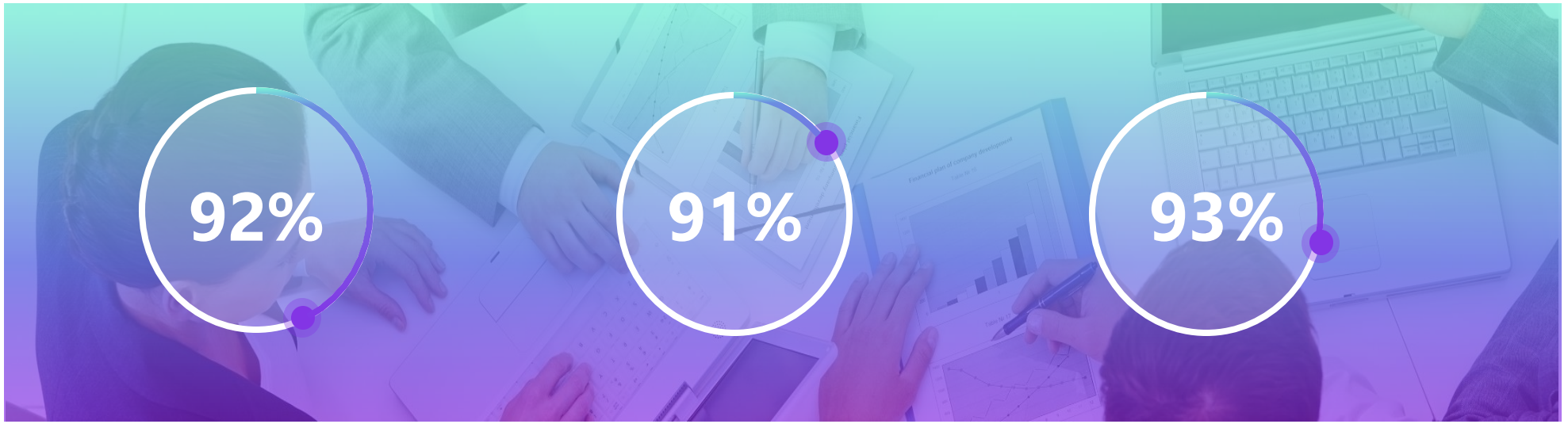


- **Managed ML Services for Evaluation:**
  AWS Sagemaker
  Google Cloud AI Platform
  Azure Machine Learning

- **Benchmarking Tests:**
  Convolutional Neural Networks(CNN)
  Transformer Models

Performance Evaluation

Model Accuracy (Average Accuracy)



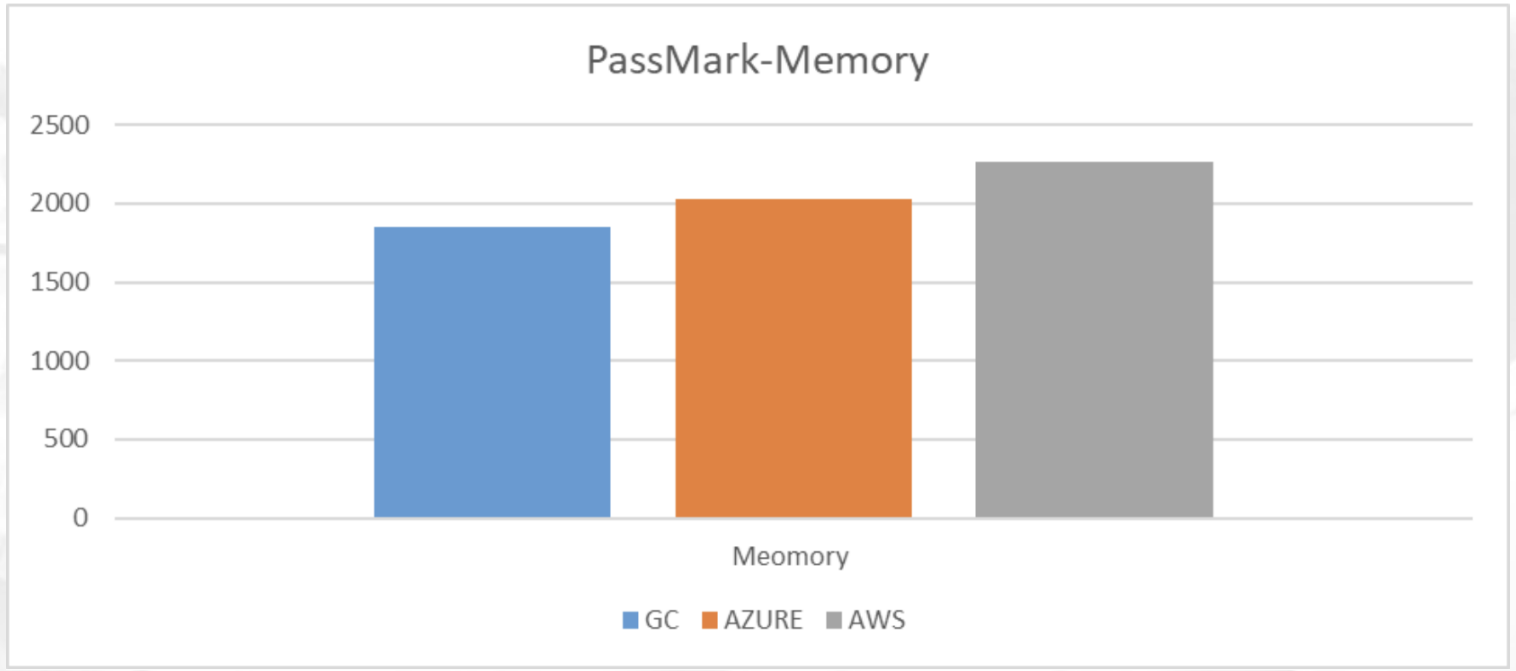
AWS	Azure	GCP
Time: 60 Minutes	Time: 65 Minutes	Time: 55 Minutes

Source: Cloud Computing Solutions for Scalable AI Model Training and Deployment

Performance on Data Access

- **Benchmarking Test:**
  PassMark Performance Test

- **Speed and efficiency in data access**

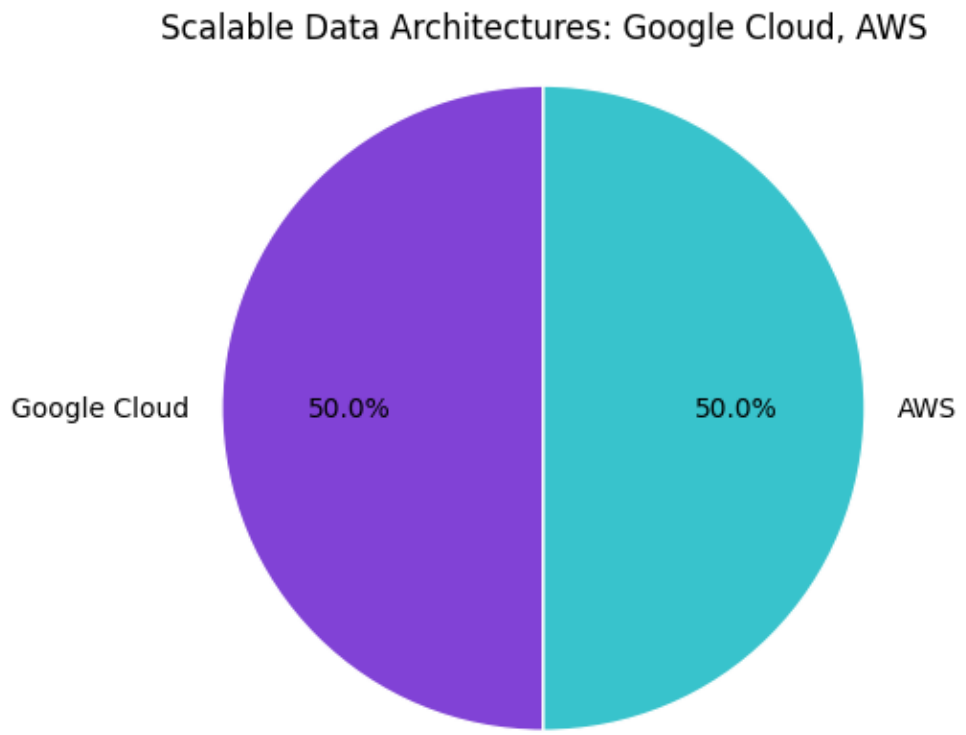


Source: [A Comparative Analysis of Cloud Computing Services: AWS, Azure, and GCP]

Performance for Scalable Data Architecture

Technologies: AI (GAN, NLP, Image Synthesis)

- **Data Process:**
  AWS Glue
  Amazon S3
  Amazon Redshift
- **Managed Service:**
  Amazon Sagemaker



- **Data Process:**
  BigQuery
  DataFlow
  Data Fusion
- **Managed Service:**
  Vertex AI

Source: [Scalable Data Architectures for Generative AI: A Comparison of AWS and Google Cloud Solutions]

Analysis of Cloud Services for Scalable AI/ML Pipelines

AMAZON AI SERVICES

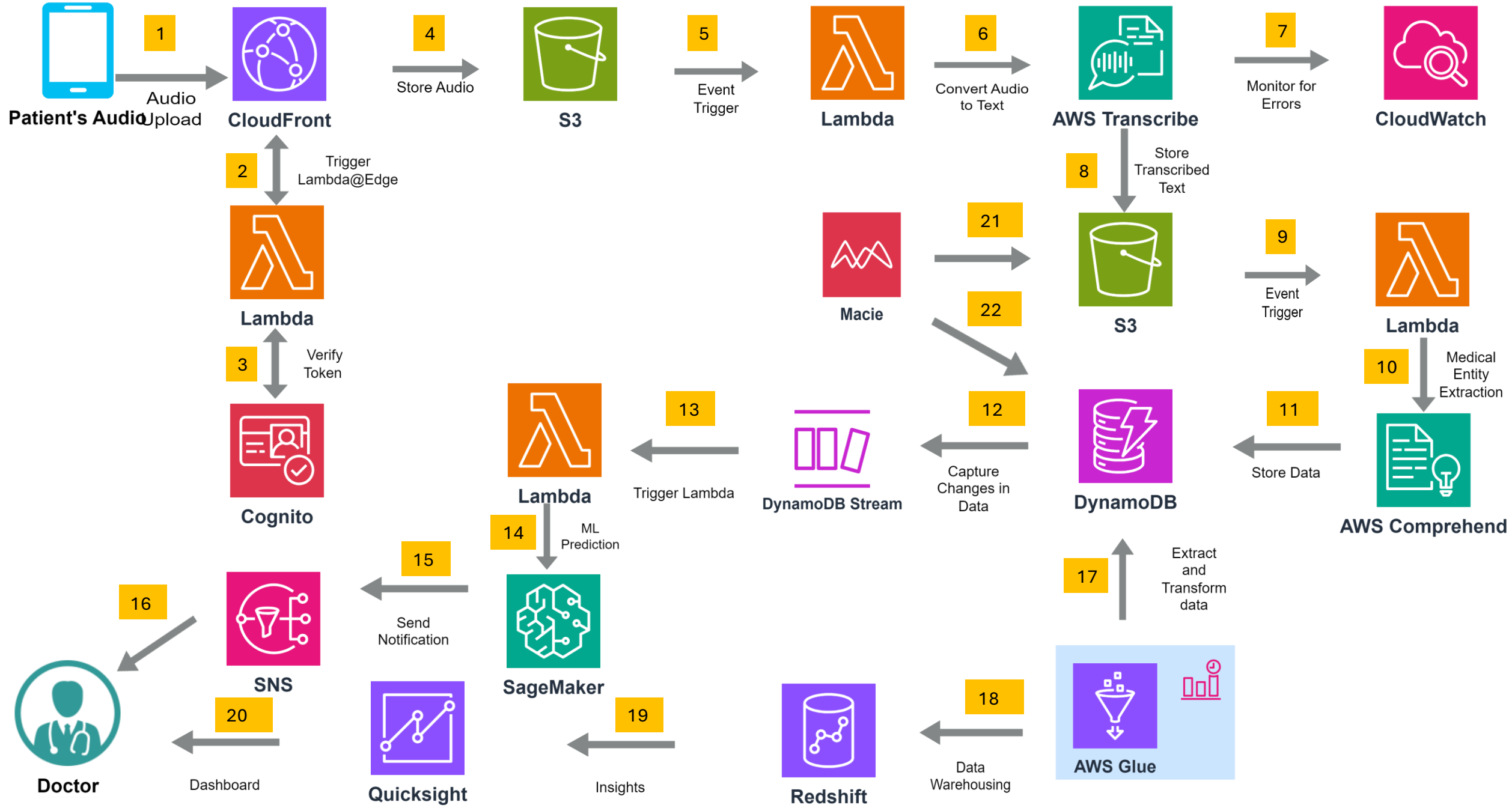
SERVERLESS ARCHITECTURE

SCALABILITY

CONTINUOUS MONITORING

CASE STUDY

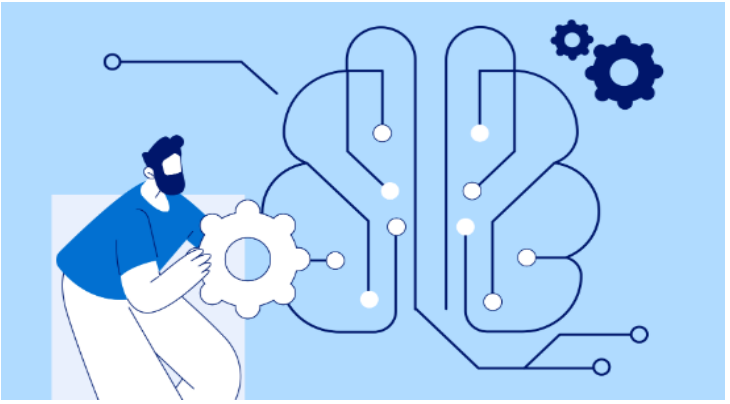
AWS Serverless Architecture for an Efficient Patient Care System



CI/CD/CT/CM in AI/ML Pipelines & **Best Practices**

Scalability Challenges in AI Pipelines

- **58%** of AI projects fail to move into full production because of challenges related to scalability.
- **Source**: https://www.gdit.com/perspectives/ai-in-full-bloom/



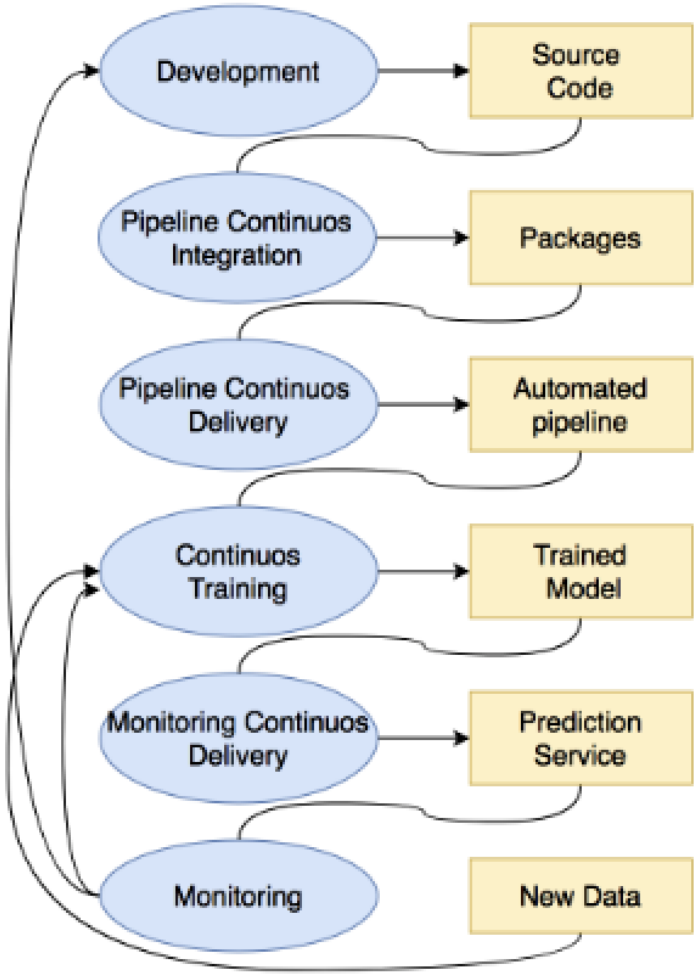
- These challenges emphasize the need for **scalable** architectures that can adapt to growing model complexity, data volumes, inference loads, and user demand.

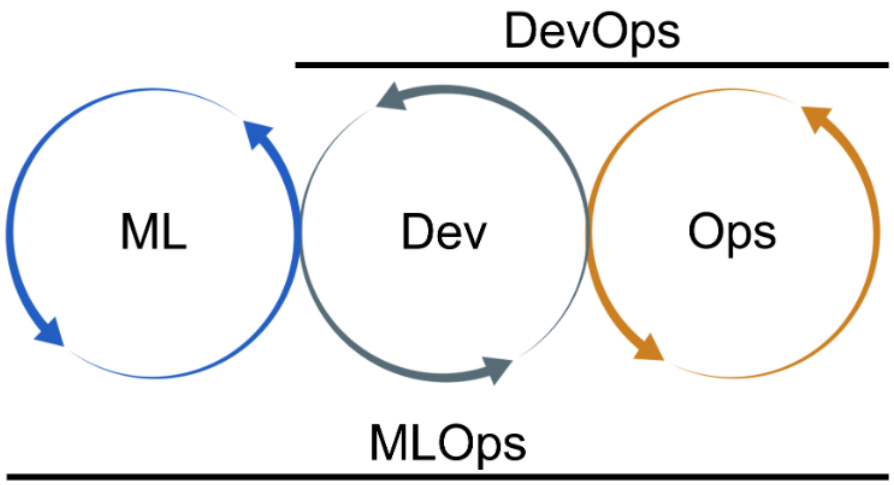
Scaling AI Pipelines: The Role of CI/CD/CT/CM for Automation and Efficiency

- "*Cloud-native **AI/ML pipelines automate** the stages of model training, validation, deployment, and post-deployment monitoring, ensuring that models remain accurate, scalable, and aligned with business objectives*"
- **Source**: *Cloud-Native AI/ML Pipelines: Best Practices for Continuous Integration, Deployment, and Monitoring in Enterprise Applications*, 2022.

How can the adoption of CI/CD/CT/CM practices ensures scalable AI Models?

CI/CD/CT/CM Process





- **Source**: *On Continuous Integration / Continuous Delivery for Automated Deployment of Machine Learning Models using MLOps, 2024*

CI/CD/CT/CM Explained

Continuous Integration (CI)

- Automates testing and merging of new code.
- Ensures model updates are integrated smoothly.
- Unit tests (AWS Codebuild), Security scanning (AWS CodeGuru), model validation tests (AWS SageMaker)

- **Tools:** Jenkins, GitLab CI, AWS CodeBuild
- **Example:** Changes to the **SageMaker** model code

Continuous Deployment (CD)

- Automates deployment of new **ML models** to production.
- **Tools:** AWS CodePipeline, GitHub Actions, Terraform.
- **Example:** Updated SageMaker deployed to production, prediction service using latest model

CI/CD/CT/CM Explained

Continuous Training (CT)

- **Retrains** the model when **new data** (e.g., patient audio) is uploaded to S3. Can be automated, manual, scheduled, on-degradation
- **Tools: SageMaker Pipelines, Kubeflow Pipelines, MLflow, Google Vertex AI, Qwak**

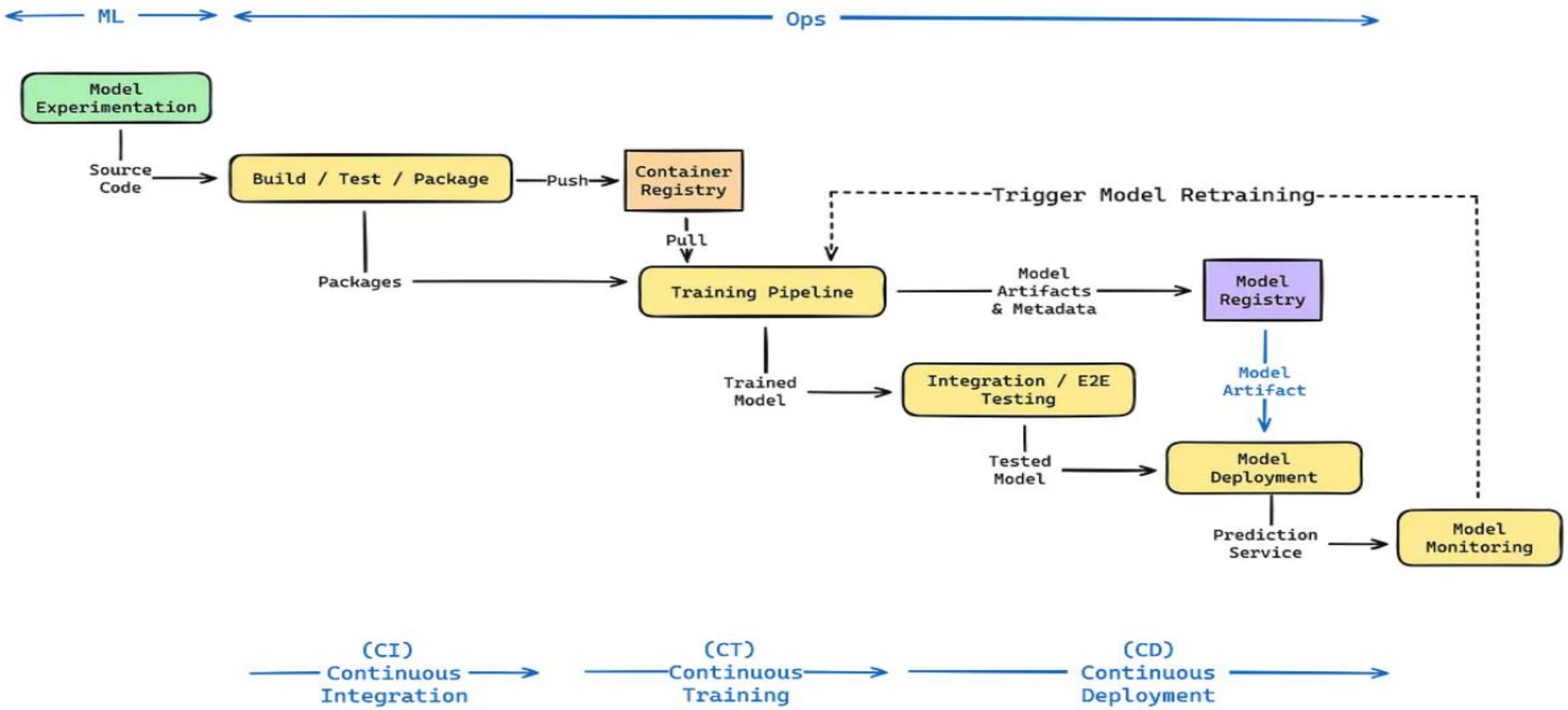
"*91% of ML Models Degrade Over Time*"
Source: https://www.fiddler.ai/blog/91-percent-of-ml-models-degrade-over-time

Continuous Monitoring (CM)

- Monitors the **performance** of the ML model during **inference**, tracking **prediction accuracy** and **response times**.

- **Tools: Prometheus**, **Grafana**, **AWS CloudWatch**
- **Example: AWS CloudWatch** monitors the **SageMaker model** as it processes patient data, triggers alarms on slow inference

CI/CD/CT/CM Flow



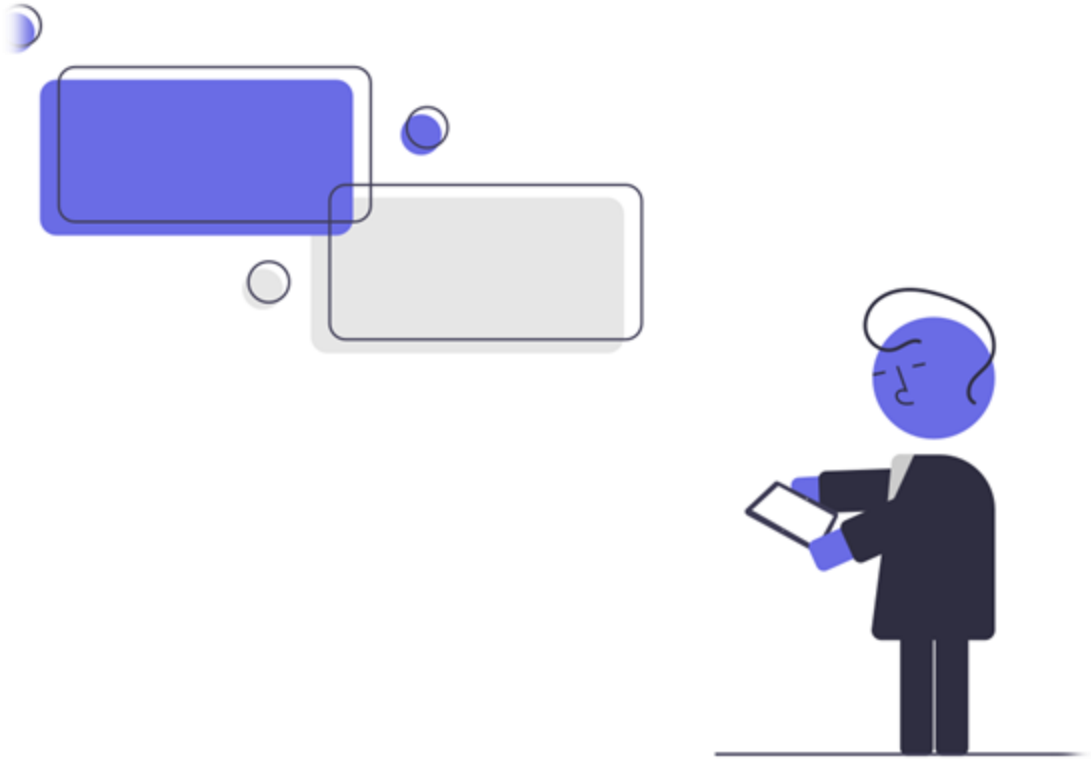
- **Source**: https://medium.com/infer-qwak/ci-cd-for-machine-learning-in-2024-best-practices-to-build-test-and-deploy-c4ad869824d2

Best Practices for CI/CD/CT/CM

- **Infrastructure as Code (IaC) :** It ensures environment is reproduceable, scalable and easily configurable across multiple stages (dev, stg)
- **Optimizing Costs with Cloud:** Use features like spot-instances (AWS EC2 Spot, GCP preemptible VM)
- **Managed Cloud ML Services:** For-instance AWS SageMaker, Google Cloud AI Platform and Azure Machine Learning to handle training, deployment, monitoring etc.
- **Auto Scaling:** Aws Auto – Scaling, GC Autoscaler or Azure Scale Sets
- **Secure Secrets Storage:** Use cloud-native secrets management tools like AWS Secrets Manager, Azure Key Vault or Google Secret Manager
- **Containerization:** Usage of AWS ECR, EKS, Google Kuberenetes Engine GKE, Azure Kubernetes Service (AKS)
- **Pre-built Cloud ML Models & Services:** for common tasks like NLP, image recognition.
  Check here: https://aws.amazon.com/marketplace/solutions/machine-learning/pre-trained-models

Challenges for AI pipelines in Cloud

- Complex Cloud Setup
- Cost Management: Autoscaling challenges
- Security Risks: Data privacy concerns
- Environmental Impact: Energy consumption and sustainability



References

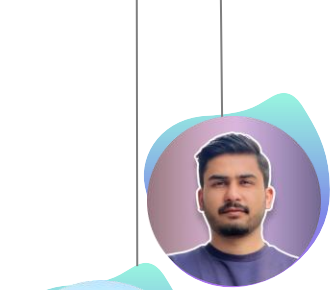
- A. Singla and T. Malhotra, “Challenges And Opportunities in Scaling AI/ML Pipelines”, J. Sci. Tech., vol. 5, no. 1, pp. 1–21, Jan. 2024.
- A. Polamarasett, "Cloud Computing Solutions for Scalable AI Model Training and Deployment”, ijaeti, vol. 1, no. 03, pp. 389–422, Sep. 2023, Accessed: Oct. 15, 2024.
- Ooijen, P. & Darzi, Erfan & Dekker, Andre. (2022). AI Technical Considerations: Data Storage, Cloud usage and AI Pipeline. 10.48550/arXiv.2201.08356.
- Bontempi, D., Nuernberg, L., Pai, S., Krishnaswamy, D., Thiriveedhi, V., Hosny, A., Mak, R. H., Farahani, K., Kikinis, R., Fedorov, A., & Aerts, H. J. W. L. (2024). End-to-end reproducible AI pipelines in radiology using the cloud. Nature Communications, 15(1). https://doi.org/10.1038/s41467-024-51202-2
- Pentyala, D. (2024, June 29). Scalable Data Pipelines in Cloud Computing: Optimizing AI workflows for Real-Time Processing. https://ijaeti.com/index.php/Journal/article/view/517
- Walia, K. (2024, April 1). Scalable AI Models through Cloud Infrastructure. https://www.espjournals.org/IJACT/ijact-v2i2p101
- Paul, Deloitte (June 2022). Cloud-Native AI/ML Pipelines: Best Practices for Continuous Integration, Deployment, and Monitoring in Enterprise Applications. https://thesciencebrigade.com/JAIR/article/view/369/349
- Garg, Pundir (Match 2022). On Continuous Integration / Continuous Delivery for Automated Deployment of Machine Learning Models using MLOps https://ar5iv.labs.arxiv.org/html/2202.03541

Our Team around the world

Abdul Wahab
Saad Abdullah

Ishara Galbokka Hewage

Tirthendu Chakravorty
Lameya Islam



Thank You

Questions are Welcomed!