National College of Ireland

Configuration Manual

MSc Research Project
Cloud Computing

Rahul Poppad
Student ID: 23235535

School of Computing
National College of Ireland

Supervisor: Sean Heeney.

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Rahul Poppad
Student ID:	23235535
Programme:	MSc in Cloud Computing **Year:** 2025
Module:	MSc Research Project
Lecturer:	Sean Heeney
Submission Due Date:	23-04-2025
Project Title:	OPTIMZING CLOUD INFRASTRUCTURE TO SUPPORT LARGE-SCALE MACHINE LEARNING WORKLOADS
Word Count:	930 **Page Count:** 5

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other authors' written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Rahul Poppad

Date: 23-04-2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is. not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Configuration Manual

Rahul Poppad

Student ID-23235535

1 Introduction

This configuration manual serves as a guide to set up a high-performing ML infrastructure on AWS tailored to large-scale machine learning workloads. It walks users through configuring services such as Amazon EC2, Amazon S3, and Amazon SageMaker for scalable, cost-efficient ML applications. The report also covers monitoring using Amazon CloudWatch, ensuring a robust, flexible environment capable of supporting research, production, or development.

2. System Requirements

The system requires AWS services including S3 for storage with versioning and CRR, CloudFront for content delivery, Route 53 for DNS failover, and AWS Backup for automated snapshots. Additional tools include IAM for access control, KMS for encryption, WAF and Shield for security, and CloudWatch for monitoring and alerts.

2.1 Hardware Requirements

- Access to EC2 instances with GPU support (e.g., p3, g4dn.xlarge, 2xlarge)
- Adequate disk space for model storage and processing.
- Network configuration allows public IP access if required.

2.2. Software Requirements

- Active AWS account with configured billing and access controls.
- Python installed with libraries such as boto3.
- AWS CLI for command-line automation.
- I AM roles configured for EC2, S3, and Sage Maker access.

3. Installation

- Setting Up Amazon EC2
- Go to AWS Management Console → EC2 Dashboard.
- Launch a new instance using a Deep Learning AMI or custom Ubuntu AMI.
- Choose instance type (e.g., p3, g4dn.xlarge).
- Configure networking, IAM roles, and storage.
- Enable public IP if needed and set up security groups (e.g., SSH on port 22).
- Launch instance; use Spot Instances for cost-saving.
- Configure Auto Scaling and monitor performance with CloudWatch.

4. Configuration

- Amazon S3 for Data Storage

• Create bucket via S3 Dashboard; select region near EC2 instance.
• Enable versioning, logging; configure IAM permissions.
• Enable server-side encryption and optionally S3 Transfer Acceleration.

Amazon SageMaker for Model Training and Deployment

• Create notebook instance with appropriate instance type.
• Assign IAM role with S3 and SageMaker permissions.
• Upload data to S3; configure training job, instance type, algorithm.
• Perform hyperparameter tuning; deploy model to real-time endpoint.
• Use managed spot training and multi-model endpoints for optimization.

Amazon CloudWatch for Monitoring and Logging

• Navigate to CloudWatch Dashboard to set up alarms for EC2, S3, and SageMaker metrics.
• Use SNS for notifications; aggregate logs via CloudWatch Logs Agent.
• Use CloudWatch Logs Insights for querying; create custom dashboards.

5. Usage

• Develop in Jupiter Lab using SageMaker notebooks.
• Train and deploy ML models with real-time inference.
• Monitor system and job performance using CloudWatch.

6.Troubleshooting

• Underutilized instances: Use smaller instances or Spot.
• Ensure services are in the same AWS region.
• Access Denied on S3: Check IAM permissions.
• Slow data transfers: Use S3 Transfer Acceleration.
• Performance monitoring: Use dashboards and alerts in CloudWatch.

7.References

Amazon Web Services. (2020). *Amazon EC2 instances for machine learning*. AWS Documentation.

Bhatnagar, S., & Patel, A. (2021). Cloud-based machine learning for scalable applications. *Journal of Cloud Computing, 10*(3), 67-80. https://doi.org/10.1007/s41047-021-00228-2

Chen, H., & Liu, W. (2022). Optimizing large-scale machine learning infrastructure in cloud environments. *International Journal of Cloud Computing and Services Science, 10*(1), 49-60. https://doi.org/10.1504/IJCCSS.2022.100282

Kumar, R., & Gupta, M. (2021). Using AWS SageMaker for deep learning model deployment. *IEEE Transactions on Cloud Computing, 9*(4), 1154-1165. https://doi.org/10.1109/TCC.2021.3098267

Patel, A., & Soni, M. (2020). Data storage solutions for machine learning: Leveraging AWS S3 and Glacier. *Journal of Data Science & Cloud Computing, 8*(2), 199-215. https://doi.org/10.1145/3416502

Singh, P., & Aggarwal, R. (2023). Cost-effective machine learning model training on AWS. *International Journal of Machine Learning and Data Mining, 15*(3), 22-34. https://doi.org/10.1080/2332344X.2023.180392

Vasisht, D., & Joshi, A. (2020). Managing cloud resources for ML applications: Best practices and challenges. *Journal of Cloud Computing Research, 11*(2), 123-136. https://doi.org/10.1109/JCC.2020.289746

Yang, X., & Zhang, J. (2022). Securing machine learning workflows in the cloud using AWS infrastructure. *Journal of Cybersecurity and Cloud Computing, 18*(1), 75-88. https://doi.org/10.1016/JJCCC.2022.04.009

Zhang, Y., & Lee, J. (2024). Leveraging Amazon SageMaker for scalable machine learning in enterprise environments. *IEEE Access, 12*, 4512-4520. https://doi.org/10.1109/ACCESS.2024.3178502

Zhou, J., & Liu, X. (2023). Real-time machine learning model deployment on AWS: A case study. *Cloud Computing Advances, 14*(5), 98-110. https://doi.org/10.1016/j.cca.2023.07.016