Configuration Manual

MSc Research Project
Cloud Computing

Priyanka Joseph
Student ID: x22114327@student.ncirl.ie

School of Computing
National College of Ireland

Supervisor: Ahmed Makki

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: ……. PRIYANKA JOSEPH……

Student ID: ……x22114327@studnet.ncirl.ie……

Programme: …Cloud Computing……

Year: ……2023-2024……..

Module: ……MSc Research Project……

Lecturer: … Ahmed Makki……

Submission Due Date: ……14/12/2023……

Project Title: …… Enhancing Water Use Data Analysis in Cloud Computing Environments through Parallel Processing Optimization……

Word Count: ……526…… **Page Count:** 8……

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

author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: ……Priyanka Joseph……

Date: ……14/12/2023……

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

Priyanka Joseph
Student ID: x22114327@student.ncirl.ie

1 Introduction

This Configuration Manual covers all local and AWS Sagemaker prerequisites for repeating the study and its outcomes. This document covers local run, dataset source, Python ML packages, AWS Sagemaker notebook instance environment configuration, and pyspark implementation.

2 Local Compute Configuration for code development

Hardware Configuration for the local run:

- Processor: Intel 11th Gen Core i7-1165G7 @2.4 GHz
- RAM: 16 GB DDR4 RAM 3200MHz
- Storage (SSD): 512GB
- Operating System: Windows 10, 64-bit

Software Packages for the local run:

- Python 3.11
- Anaconda Navigator 2.3.2
- PyCharm IDE Community Edition 2021.3
- Jupyter Lab

3 Data Science Environment and Packages

- Pip/Conda to install packages and dependencies
- Java 8 environment: Install using
  `apt install -y openjdk-8-jdk openjdk-8-jre`
- Install pyspark `pip install pyspark`
- Other packages:
  - Pandas
  - Numpy
  - Matplotlib
  - Seaborn

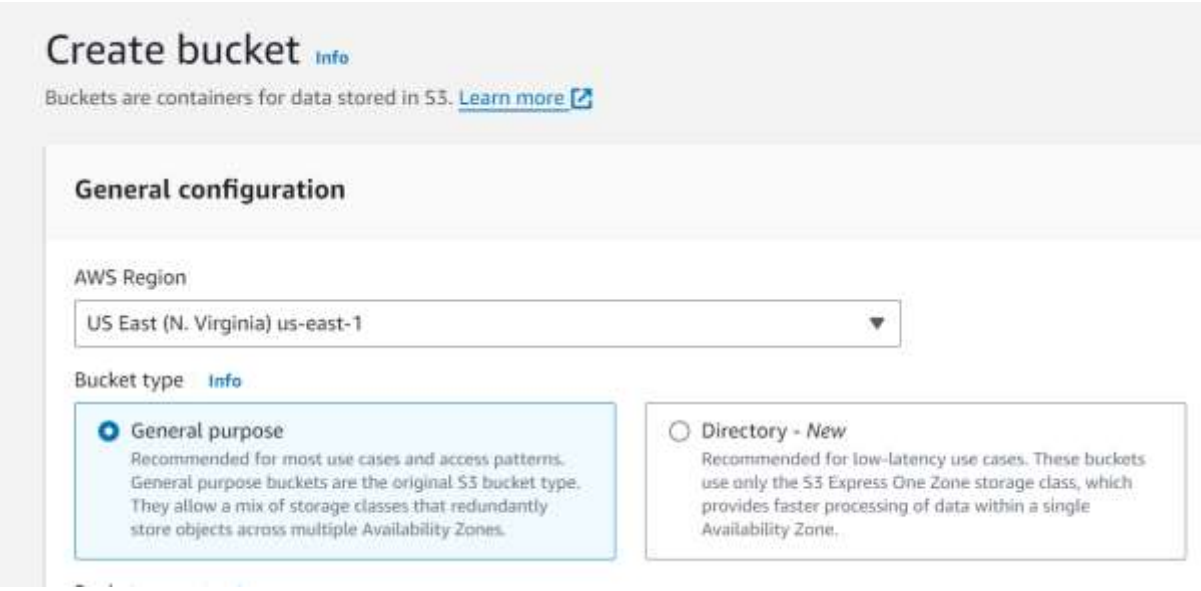
4 Dataset

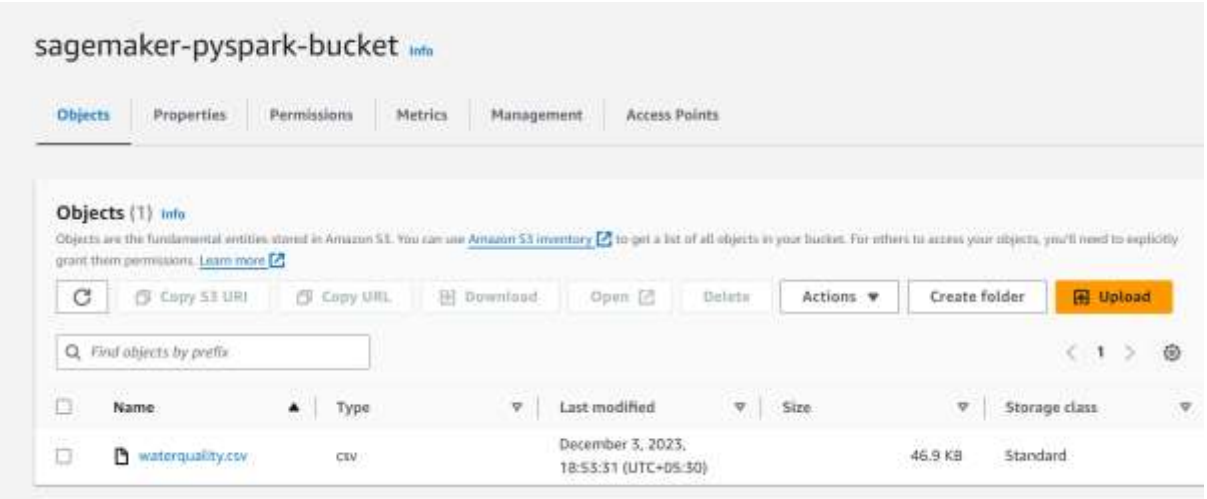
The Indian Central Pollution Control Board (CPCB) website, specifically the National Water Quality Monitoring Network Programme (NWMP), provided the data. We collected this data to assess water quality in diverse water resources and avoid and manage water contamination. Surface and groundwater are monitored by 4484 stations in 28 states and 8 territories. Regular monitoring occurs monthly, quarterly, semi-annually, and annually. Water samples are analyzed for 9 key characteristics and compared to CPCB's optimal water quality guidelines. Dissolved oxygen (DO), pH, conductivity, biological oxygen demand (BOD), nitrate, fecal coliform, and total coliform are in the Kaggle dataset. This information ensures potable water quality.

Data Source: https://cpcb.nic.in/nwmp-data/
Dataset Download Link: https://www.kaggle.com/datasets/akkshaysr/nwmp-water-quality-data-for-indian-lakes/data

5 AWS Sagemaker Configuration

5.1 S3 bucket creation and uploading the waterquality.csv file.





5.2 IAM permissions for AWS Glue

AWS Glue IAM permissions are configured as follows.

1. Establish an IAM policy for AWS Glue service, enabling access to S3 storage, EC2 instances, and CloudWatch metrics.

Set up an IAM role for AWS Glue to grant it authority to request additional services on your behalf. Amazon S3 can be used for all AWS Glue sources, targets, scripts, and temporary folders.

3. Attach GlueConsoleAccessPolicy, GlueAWSGlueConsoleSageMakerNotebookFullAccess, and AWSCloudFormationReadOnlyAccess policies to users or groups accessing AWS Glue.
4. Create an IAM policy for notebook servers. This policy authorizes Amazon S3 activities to manage your account's resources for AWS Glue's role.
5. Create an IAM role for notebook servers to enable AWS Glue to request other services on your behalf. Amazon S3 can be used for all AWS Glue sources, targets, scripts, and temporary folders.
6. Finally, establish an IAM role for SageMaker notebooks.

5.3 Configuring a notebook instance in AWS Sagemaker

Step 1: Navigate to Amazon SageMaker

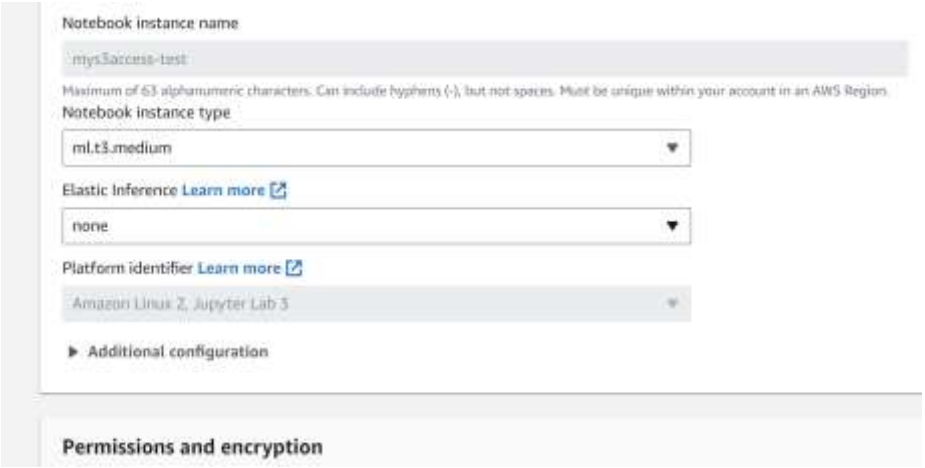
1. In the AWS Management Console, navigate to the Amazon SageMaker service.

Step 2: Create a Notebook Instance

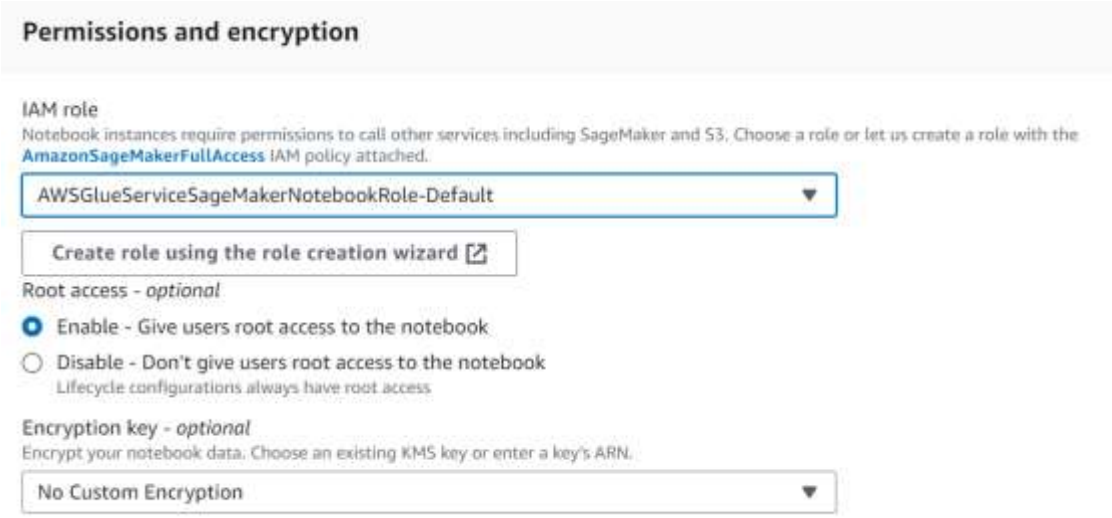
1. In the SageMaker console, click on Notebook instances in the left navigation pane.
2. Click on the Create notebook instance button.

Step 3: Configure Notebook Instance Settings

1. Provide a Notebook instance name.
2. Choose an IAM role AWSGlueServiceSageMakerNotebookRole-Default that has the necessary permissions for SageMaker.
3. Choose an instance type ml.t3.medium.
4. Click on Create

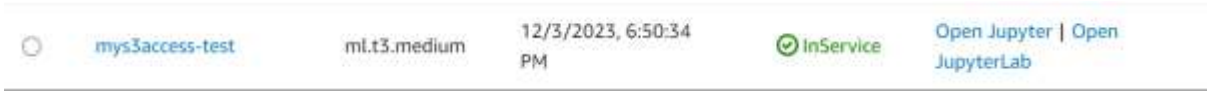


Step 4: Configure IAM role with AWSGlueServiceSageMakerNotebookRole-Default selected.

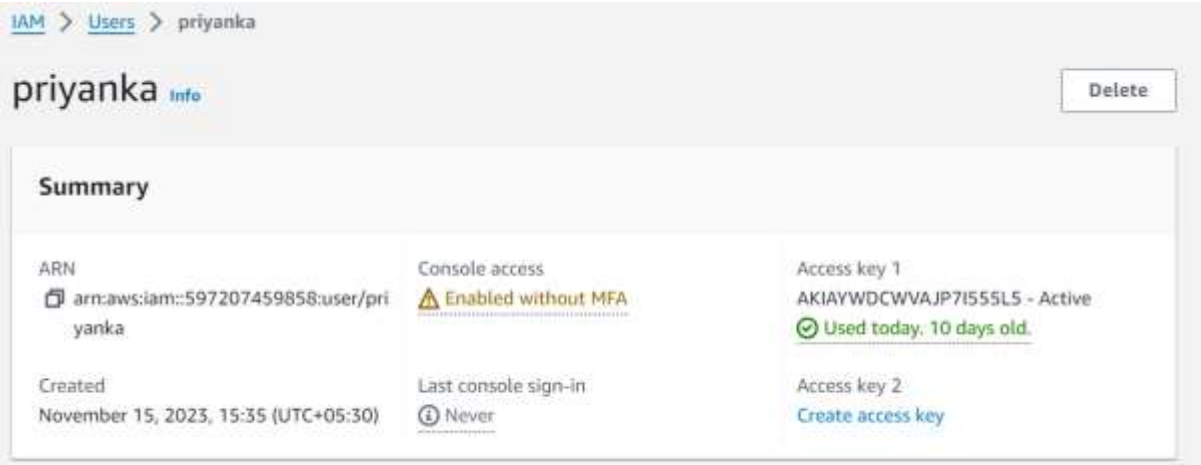


Step 5: Click *Create* button to enable notebook instance creation

Step 6. This will create the notebook instance and put it in service.



5.4 IAM Credentials Access Key creation



6 Running Notebook Instance

The instance creates a Jupyter lab notebook lab for code development and deployment.