

# 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**..... PRIYANKA  
**Name:** JOSEPH.....  
.....

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

**Progra** ...Cloud **Year** .....2023-  
**mme:** Computing..... : 2024.....  
..... .....

**Module:**.....MSc Research  
Project.....  
.....

**Lecture** ... Ahmed  
**r:** Makki.....  
.....

**Submis**  
**sion** .....14/12/2023.....  
**Due** .....  
**Date:**

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

**Word** .....526..... **Page Count:**  
**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.

**Signatu** .....Priyanka

**re:** 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)	<input type="checkbox"/>
<b>Attach a Moodle submission receipt of the online project submission,</b> to each project (including multiple copies).	<input type="checkbox"/>
<b>You must ensure that you retain a HARD COPY of the project,</b> both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	<input type="checkbox"/>

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

<b>Office Use Only</b>	
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 11<sup>th</sup> 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

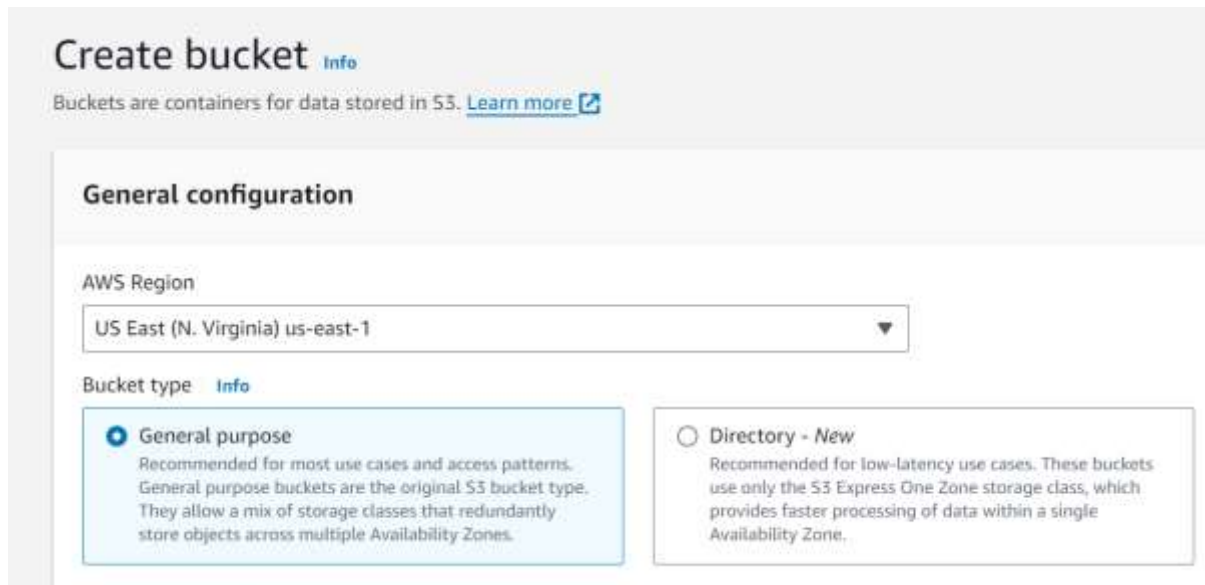
The Indian Central Pollution Control Board (CPCB) website, specifically the National Water Quality Monitoring Network Programme (NWMP), provided the data. We collected this data to assess water quality in diverse water resources and avoid and manage water contamination. Surface and groundwater are monitored 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.



**Create bucket** [Info](#)

Buckets are containers for data stored in S3. [Learn more](#) [↗](#)

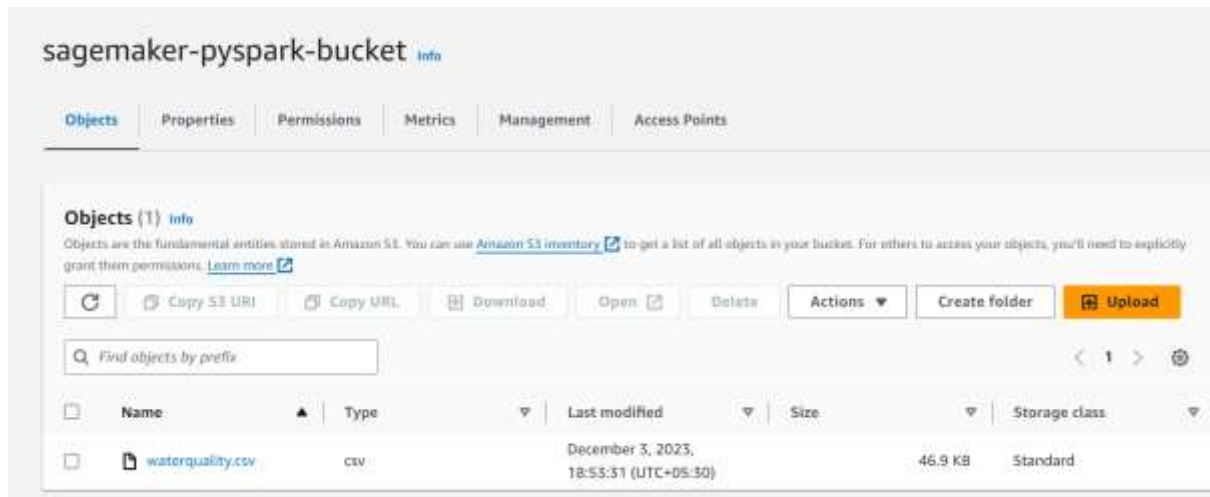
### General configuration

AWS Region

US East (N. Virginia) us-east-1 ▼

Bucket type [Info](#)

- General purpose**  
Recommended for most use cases and access patterns. General purpose buckets are the original S3 bucket type. They allow a mix of storage classes that redundantly store objects across multiple Availability Zones.
- Directory - New**  
Recommended for low-latency use cases. These buckets use only the S3 Express One Zone storage class, which provides faster processing of data within a single Availability Zone.



## 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

**Notebook instance name**  
 mys3access-test  
Maximum of 63 alphanumeric characters. Can include hyphens (-), but not spaces. Must be unique within your account in an AWS Region.

**Notebook instance type**  
 ml.t3.medium

**Elastic Inference** [Learn more](#)  
 none

**Platform identifier** [Learn more](#)  
 Amazon Linux 2, Jupyter Lab 3

▶ **Additional configuration**

**Permissions and encryption**

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

**Permissions and encryption**

**IAM role**  
 Notebook instances require permissions to call other services including SageMaker and S3. Choose a role or let us create a role with the `AmazonSageMakerFullAccess` IAM policy attached.

AWSGlueServiceSageMakerNotebookRole-Default

[Create role using the role creation wizard](#)

**Root access - optional**

**Enable** - Give users root access to the notebook

**Disable** - Don't give users root access to the notebook  
Lifecycle configurations always have root access

**Encryption key - optional**  
 Encrypt your notebook data. Choose an existing KMS key or enter a key's ARN.

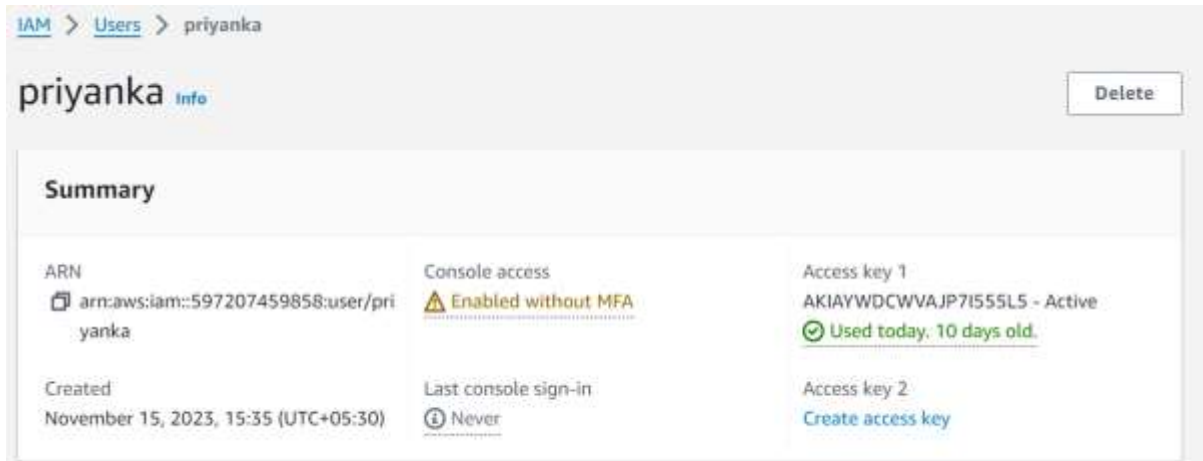
No Custom Encryption

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

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

<input type="radio"/>	mys3access-test	ml.t3.medium	12/3/2023, 6:50:34 PM	<span style="color: green;">✔</span> InService	<a href="#">Open Jupyter</a>   <a href="#">Open JupyterLab</a>
-----------------------	-----------------	--------------	-----------------------	--	--

## 5.4 IAM Credentials Access Key creation

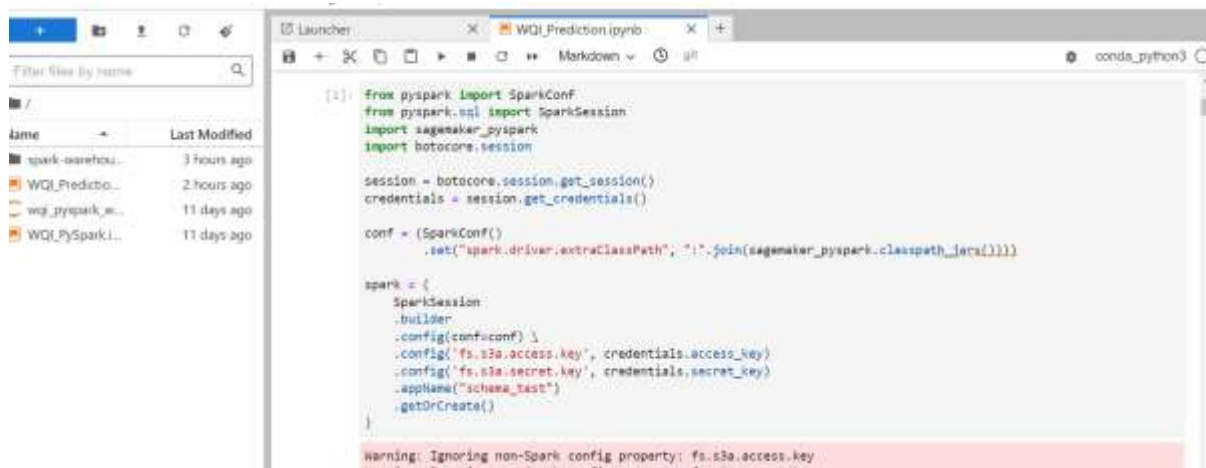


The screenshot shows the AWS IAM console for user 'priyanka'. The 'Summary' section displays the following information:

Property	Value
ARN	arn:aws:iam::597207459858:user/priyanka
Console access	Enabled without MFA
Access key 1	AKIAYWDCWVAJP7I555L5 - Active Used today, 10 days old.
Access key 2	Create access key
Created	November 15, 2023, 15:35 (UTC+05:30)
Last console sign-in	Never

## 6 Running Notebook Instance

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



The screenshot shows a Jupyter Lab notebook with the following Python code:

```
[1]: from pyspark import SparkConf
from pyspark.sql import SparkSession
import sagemaker_pyspark
import boto3.session

session = boto3.session.Session()
credentials = session.get_credentials()

conf = (SparkConf()
        .set("spark.driver.extraClassPath", "!".join(sagemaker_pyspark.classpath_libs)))

spark = (
    SparkSession
    .builder
    .config(conf=conf) \
    .config('fs.s3a.access.key', credentials.access_key)
    .config('fs.s3a.secret.key', credentials.secret_key)
    .appName("schema_test")
    .getOrCreate()
)
```

A warning message is displayed at the bottom: "Warning: Ignoring non-Spark config property: fs.s3a.access.key".