

Configuration Manual

MSc Research Project Cloud Computing

Bhuvan Prashanth Student ID: 22163654

School of Computing National College of Ireland

Supervisor: Dr. Rashid Mijumbi

National College of Ireland Project Submission Sheet School of Computing



Student Name:	Bhuvan Prashanth
Student ID:	22163654
Programme:	Cloud Computing
Year:	2023
Module:	MSc Research Project
Supervisor:	Dr. Rashid Mijumbi
Submission Due Date:	14/12/2023
Project Title:	Efficient Real-Time Data Deduplication Techniques for Im-
	proving Data Quality in Urban Taxi Trip Streams
Word Count:	425
Page Count:	6

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.

<u>ALL</u> 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:	Bhuvan Prashanth
Date:	14th December 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

Bhuvan Prashanth 22163654

1 Introduction

Outlined here are the steps for configuring the Notebook file on AWS SageMaker.

1.1 Requirement

Before commencing the setup process, verify that the following requirements have been fulfilled:

- Possession of data in CSV format.
- An active AWS account.
- Access to SageMaker with Compute-Optimized instances.
- Proficiency in Boto 3 and Python.
- Install Pycharm , NymPy, Dask, Matlib libraries
- Optionally, contemplate the use of VScode for future tasks.

1.2 Implementation

1. Log in to your AWS Account and Go to AWS S3 buckets

aws	Services Q Search	[^	It+S] D ♦ Ø Ø Ireland ▼ bhuvanp9_6 ▼
=	Console Home Info		Reset to default layout + Add widgets
	:: Recently visited Info	:	# Applications (0) Info Create application : Region: Europe (Ireland)
		< 1 2 >	
	53	Billing and Cost Management	eu-west-1 (Current Region) V Q. Find applications
	Amazon SageMaker	Service Quotas	< 1 >
	₽ EC2	CodeDeploy	Name ▲ Description ▼ Region ▼ Originating a
	ET IAM	6 Simple Notification Service	
	Support	CodeBuild	No applications Get started by creating an application.
	AWS Organizations	Seanstalk	Create application
	View	all services //	Go to myApplications

Figure 1: AWS Console Page

2. Next step Create S3 Bucket for storing csv files.

Services Q Search			[Alt+S]		Ð	\$ @	۲	Global 🔻	
azon S3 ×	Amazon	1 53							
ess Grants New	► A st	ccount snapshot torage lens provides visibility into storage usa	ge and activity trends. Learn more 🔀			View St	orage Le	ns dashboard	8
ss Points ct Lambda Access Points i-Region Access Points	Gene	eral purpose buckets Directory b	puckets						
h Operations Access Analyzer for S3	Gen	teral purpose buckets (3) Info	more 🔀	C Copy ARN	Empty	Delet	e	Create bucke	t
h Operations Access Analyzer for S3 k Public Access settings this account	Gen Bucke	teral purpose buckets (3) Info Its are containers for data stored in S3. Learn Find buckets by name	more 🗗	C Copy ARN	Empty	Delet	e	Create bucke	•
h Operations Access Analyzer for S3 k Public Access settings his account age Lens	Gen Bucke	eral purpose buckets (3) Info ts are containers for data stored in 53. Learn Find buckets by nome	more 🖸	C Copy ARN	Empty v	Delet	e e	Create bucke	• ©
h Operations Access Analyzer for S3 R Public Access settings his account age Lens boards and ans resume	Gen Bucke	eral purpose buckets (3) Info ts are containers for data stored in 53. <u>Learn</u> Find buckets by name Name nyuberbucketcsv	AllyS Region	C Copy ARN Acces Objects can be public	Empty	Creation Decemb (UTC+05	e date er 7, 2023 :30)	Create bucke	• ©
h Operations Access Analyzer for 53 Public Access settings his account age Lens boards age Lens groups Organizations settings	Gen Bucke	Pertal purpose buckets (3) Infe ets are containers for data stored in 53. Learn Find buckets by nome Name myuberbucketcov sagemaker-es-west-1- 04565477848667	More C AWS Region Europe (Ireland) eu-west-1 Europe (Ireland) eu-west-1	C Copy ARN	Empty v	Creation Decemb (UTC+05 Decemb (UTC+05	e date er 7, 2023 :30) er 7, 2023 :30)	Create bucke	ŧ ⊚

Figure 2: AWS S3 Bucket

3. Upload the CSV files to the s3 bucket.

₹₩\$ III Services Q. Search	[Alt+5]	D.	÷	Ø	۲	Global v	bhuve	np9_6 v
Amazon S3 ×								٩
Buckets Access Grants New Access Points Object Lamdba Access Points Multi-Region Access Points Butch Operations	Objects (Properties Permitteent Mattices Materials Access Points Objects (7) not Objects	licitly grant ad	them per	missions.)	aarn mor			0
IAM Access Analyzer for S3	Name ▲ Type ▼ Last modified ▼ Size ▼ Storage class							
Block Public Access settings for this account	December 7, 2023, Bogeta.cv/ December 7, 2023, 01:46/32 (UTC=06:30) 1.5 MB Standard							
Storage Lens Dashboards	Travel_Times - Boston.csv December 7, 2023, 01:46/38 (UTC+05:30) 967.3 KB Standard							
Storage Lens groups AWS Organizations settings	Dimest Times - December 7, 2023, Johannedzorg and cov 01-46-40 (UTC+06-30) 36.4 KB Standard							
Feature spotlight	Travel_Times - December 7, 2023, 331.8 KB Standard Manila.csv 01.46x43 (UTC+05.30) 331.8 KB Standard							
	Travel_Times - Paris.cov December 7, 2023, 01:46:46 (UTC+05:30) 324.9 KB Standard							
AWS Marketplace for S3	Travel_Times - Sydney.cov December 7, 2023, 01:46:51 (UTC+05:30) 590.0 KB Standard							
	D Travel_Times - Washington DC.cov December 7, 2023, 0147/25 (UTC+05:30) 4.1 MB Standard							

Figure 3: Adding CSV files to S3 Bucket

4. Open IAM in new tab and configure user and its roles with s3 Full AccessPermision

uWS III Services Q lam	× •	
Amazon S3	Search results for 'lam' Try searching with longer queries for more relevant resu	ults ket
Buckets Services (Access Grants New Access Points Document Object Lambda Access	(21) Services (New) tation (48,933) I I AM 12 Manana access to AWS resources.	See all 11 results a 3 is stored in a bu to 53, you'll need to 53, you'll need to bijects will be st
Multi-Region Access P Knowledg Batch Operations Marketpla IAM Access Analyzer f Blogs (1,7 Events (12	permease (ssi) Image and the state of the s	accounts and cloud applications
Block Public Access se for this account	(2) Resource Access Manager 🔅 Share AWS resources with other accounts or AW	/S Organizations
 Storage Lens Dashboards Storage Lens groups 	AWS App Mesh ☆ Easily monitor and control microservices	re no minimum fe
AWS Organizations se	Features	See all 21 results >
Feature spotlight 👩	Groups	onthly bill using the tor C

Figure 4: IAM Role

5. Setup Policies for User

aws Services Q Search		[Alt+S]	🗘 🔞 🙆 Global 🔻 bhuvanp9_6 🔻
Identity and Access × Management (IAM)	IAM > Policies		0
Q. Search IAM	Policies (1/1165) Info A policy is an object in AWS that defines permission	C Actions V	Delete Create policy
Dashboard	Q s3	Filter by Type X All types 12 matches	< 1 > ©
▼ Access management	Policy name	Type 🗢 Used as 🗢	Description
User groups	AmazonDMSRedsh	AWS managed None	Provides access to manage \$3 settings
Users		AWS managed None	Provides full access to all buckets via t
Roles		Awa manageo none	Howdes four access to all buckets via t
Policies	O	AWS managed None	Provides AWS Lambda functions perm
Identity providers	O I AmazonS3Outpost	AWS managed None	Provides full access to Amazon S3 on .
Account settings	○	AWS managed None	Provides read only access to Amazon S
Access reports	O 🚺 AmazonS3ReadOn	AWS managed None	Provides read only access to all bucke
Access Analyzer		-	
External access	O ➡ ➡ AWSBackupService	AWS managed None	Policy containing permissions necessar

Figure 5: Set Policies to S3 bucket

6. Create notebook instance to upload the .ipynb and dataset files as shown

💭 jupyter	Open JupyterLab Quit Logout
Files Running Clusters SageMaker Examples Conda	
Select items to perform actions on them.	Upload New - 2
	Name 🔶 Last Modified File size
C miller107-travel-time-uber-movement	4 days ago
Deduplication and Parallel processing uber movement data_with HIBD.ipynb	Running 9 minutes ago 3.61 MB
Without Hased Based Indexing.ipynb	Running 2 days ago 502 kB
Travel_Times_Boston.csv	10 hours ago 1.01 MB

Figure 6: Jpyter Notebook

2 Python Libraries

1. Pip Install all the dependencies



Figure 7: Importing Boto3 library

2. Let us import the necessary libraries.

```
import pandas as pd
import matplotlib.pyplot as plt
from shapely.geometry import Polygon
from geopandas import GeoDataFrame
```

Figure 8: Import the python library

- 3. Install running the command "Pip install boto3 pandas shapely geopandas"
- 4. Run the HIBD notebook after configuring your bucket



Figure 9: Configuring S3 credentails in the HIBD code

5. Output related to Performance Metrics.



Figure 10: Output for HIBD Performace Metrics

6. Repeat the same procedure for the second notebook "Dask Parallel Process".



Figure 11: Dask Parallel Process Code

7. Output for Mean travel time for Dask



Figure 12: Dask Parallel Process Code

8. Performance Metrics Output Graphs for HIBD and Dask



Figure 13: Deduplication Accuracy for HIBD & Dask



Figure 14: Performance Speed for HIBD & Dask



Figure 15: Resource consumption for HIBD & Dask

3 Dependency Installations

- 1. Install "pip install package-name" command
- 2. Install "Pip Install pandas"
- 3. Install "pip install matplotlib"
- 4. Install "Pip install geopandas"