

# **Configuration Manual**

MSc Research Project MSc Cloud Computing

Sumit Kumar Sahoo Student ID: 21154589

School of Computing National College of Ireland

Supervisor: S

Sean Heeney

### National College of Ireland



### **MSc Project Submission Sheet**

	School of Computing		
Student Name: Student ID:	Sumit Kumar Sahoo 21164967		
Programme:	Msc in Cloud Computing	Year:	2022-2023
Module:	Msc Research Project		
Supervisor: Submission Due	Sean Heeney		
Date:	01/01/2023		
Project Title:	Open-source ETL Framework using Big Data tools & Orchestration on AWS Cloud Platform	a m	

### Word Count: 1091 Lines Page Count: 18

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:

Date:

01/01/2023

### 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**

Sumit Kumar Sahoo Student ID: 21164967

# **Open-source ETL Framework using Big Data tools and Orchestration on AWS Cloud Platform**

## **1** Introduction

This is a configuration manual for Setting out Tools, Software and AWS services used for Developing a Big Data ETL framework using Python, Pyspark and Apache airflow for orchestration. We have used Terraform for Infrastructure as Code. We have used S3 bucket for using it as Source and destination and Redshift For Data warehouse Data Analytics. We have used CloudTrail for Auditing and AWS price calculator for calculating Total Cost of Operation.

## 2 Tools and Softwares Used

- 1. Visual Studio Code IDE
- 2. GitBash
- 3. Python 3.8
- 4. Terraform => 0.12
- 5. PySpark
- 6. Apache Airflow 2.2
- 7. Aamazon Webservices

## **3** Software Installation

## 3.1 Visual Studio Code –

1. Download Visual Studio Code for Windows using <u>https://learn.microsoft.com/en-us/visualstudio/install/install-visual-studio?view=vs-2022</u>

## 3.2 Setup Terraform

1. Download Terraform for Windows using - <u>https://developer.hashicorp.com/terraform/downloads</u>

🕪 🗸 🚩 Terrafor	rm	Home Documentation ~ Tutorials Install Registry C Try Cloud C		Q Search #/ctrl K Q ~
< Terraform Home Install Terraform		Operating System macOS Windows Linux FreeBSD OpenBSD Solaris	🍸 1.3.6 (latest) 🗘	About Terraform Define cloud and on-prem
Getting Started		Binary download for Windows		resources in human-readable configuration files that you can version, reuse, and share.
What is Terraform? Terraform Tutorials		386 Version: 13.6	Download 🕁	Featured docs
Terraform Cloud Tutorials		AMD64 Version: 1.3.6	Download 소	Configuration Language Terraform CLI
Tutorial Library		Release information		Terraform Cloud Provider Use
Community Forum Support	ď	Changelog Version: 1.3.6	GitHub 🗗	
GitHub Terraform Registry	C C	Notes You can find the <u>SHA256 checksums for Terraform 1.3.6</u> online and you can <u>verify the checksums</u> signed using <u>HashiCorp's GPG key</u> .	s signature file which has been	Y Terraform Cloud Automate your infrastructure provisioning at any scale     The Terraform Cloud     →

- 2. Extract the .zip file and store the folder in C:\ Drive
- 3. Setup environment and user variables by Editing the Path and pasting the Path where the above folder is
  - copied.

Edit environment vari	able	×
%SystemRoot%\Sy	vstem32\Wbem	New
%SYSTEMROOT%\	System32\WindowsPowerShell\v1.0\	
%SYSTEMROOT%\	System32\OpenSSH\	Edit
%USERPROFILE%\a	anaconda3	
%USERPROFILE%\a	anaconda3\Scripts	Browse
%USERPROFILE%\a	anaconda3\Library\bin	
C:\Program Files\Ja	ava\jdk1.8.0_301\bin	Delete
C:\ProgramData\c	hocolatey\bin	Delete
C:\Program Files\D	Oocker\Docker\resources\bin	
C:\ProgramData\D	ockerDesktop\version-bin	Movellp
C:\Program Files (x	86)\Symantec\VIP Access Client\	Move op
C:\Program Files\A	mazon\AWSCLIV2\	Maria Davia
C:\Program Files (x	(86)\Bitvise SSH Client	Move Down
C:\Program Files\n	odejs\	
C:\Program Files\P	uTTY\	
C:\Program Files (x	86)\Microsoft SQL Server\160\Tools\Binn\	Edit text
C:\Program Files\N	licrosoft SQL Server\160\Tools\Binn\	
C:\Program Files\N	licrosoft SQL Server\160\DTS\Binn\	
C:\Program Files (x	86)\Microsoft SQL Server\160\DTS\Binn\	
C:\Program Files\A	zure Data Studio\bin	
C:\terraform		
	OK	Cancel

4. Goto the Visual Studio and open the Project and right click on ec.tf file -> Click on Open in integrated terminal -> cmd and run terraform plan

# 4 Deployment of EC2 which hosts Apache Airflow and S3 in us-east-1



1. Run command -> terraform plan



This will start automated deployment of :
 a. S3 Bucket – emrbuckettestsumit

← → C	n/s3/buckets?region=us-east-1 Python Airflow Conda-Issues Logging Cert	ifications 🧧 Spark ETL-upsert 🧧 Me	orningstar 📒 MyUse 💶 Java-Knowled	년 🛧 🚳 편 🔰 💥 🏚 🖬 🌒 Ige 📙 Java-Knowledge 📙 Imported
aws Services Q Search	[Alt+S]			Ç O Global ▼ x21154589-sumit-thesis
Amazon S3 ×	Amazon S3 > Buckets			
Buckets Access Points Object Lambda Access Points	Account snapshot     Storage lens provides visibility into storage usage and activity tree	nds. Learn more 🔀		View Storage Lens dashboard
Multi-Region Access Points Batch Operations Access analyzer for S3	Buckets (4) Info Buckets are containers for data stored in 53. Learn more 🖸 Q. Find buckets by name	C	Copy ARN Empty	Delete Create bucket
Block Public Access settings for	Name 🔺	AWS Region 🗢	Access Cre	ation date 🗢
Storage Lens	aws-cloudtrail-logs-505316317378-d9e24370	US East (N. Virginia) us-east-1	Bucket and objects not Dec	ember 10, 2022, 04:57:22 (UTC+00:00)
Dashboards	O aws-logs-505316317378-us-east-1	US East (N. Virginia) us-east-1	Objects can be public Dec	ember 4, 2022, 20:50:13 (UTC+00:00)
AWS Organizations settings	O emrbuckettestsumit	US East (N. Virginia) us-east-1	Objects can be public Dec	ember 12, 2022, 18:56:00 (UTC+00:00)
	x21154589emrbucket	US East (N. Virginia) us-east-1	Bucket and objects not Dec	ember 3, 2022, 20:07:50 (UTC+00:00)
Feature spotlight 3 AWS Marketplace for S3				

b. Create files as shown below

aws Services Q Search	[Alt+5]	۵.	)   \$	0	Global 🔻	x2115458	39-sumit-th
Amazon S3 ×	Amazon S3 > Buckets > emrbuckettestsumit						
Buckets	emrbuckettestsumit Info						
Access Points Object Lambda Access Points Multi-Region Access Points Batch Operations Access analyzer for S3	Objects         Properties         Permissions         Metrics         Management         Access Points           Objects (3)         Image: State St						
Block Public Access settings for this account	Objects are the fundamental entities stored in Amazon 53. You can use Amazon 53 inventory <b>C</b> to get a list of all objects in your bucket. F grant them permissions. Learn more <b>C</b> <b>C C C</b> Copy 53 URI <b>C</b> Copy URL <b>D</b> Download <b>Open C D</b> Delete <b>Action</b> <b>Q</b> . Find objects by prefix	For others	to access y	our objects a <b>te fold</b> e	s, you'll need to	Upload	0
Dashboards AWS Organizations settings	Name         ▲         Type         ▼         Last modified         ▼           ▶ x21154589_input/         Folder         -	Size	▽	Sto	rage class		▽
Feature spotlight 3	Image: science		-	-			
AWS Marketplace for S3							

c. Create EC2 c3.large in us-east-1

aws iii Services Q Search	h [	Alt+S]	▶ ♦ ⑦ N. Virginia ▼ x21154589-sumit-thesis ▼
New EC2 Experience	Instances (1/1) Info	Connect Instance	e state 🔻 Actions 🔻 Launch instances 🔻
Tett us what you think	Q. Find instance by attribute or tag (case-sensitive)		< 1 > @
EC2 Dashboard	Instance state = running X Clear filters		
EC2 Global View		-	
Events	✓ Name ♥ Instance	e ID Instance state	
Tags	web-server i-0b087	2a193e51d5f2	⊘ 2/2 checks passed No alarms + us-east-1a
Limits			•
▼ Instances			
Instances		=	
Instance Types	Instance: i-0b0872a193e51d5f2 (web-server)		© ×
Launch Templates	Details Security Networking Storage	Status cilecitis monitoring rags	
Spot Requests			
Savings Plans	▼ Instance summary Info		
Reserved Instances	Instance ID	Public IPv4 address	Private IPv4 addresses
Dedicated Hosts		E 34.137.120.45 Open address E	<b>D</b> 10001130
Scheduled Instances	IPv6 address	Instance state	Public IPv4 DNS
Capacity Reservations		Channing	
	Hostname type	Private IP DNS name (IPv4 only)	
▼ Images	IP name: Ip-10-0-1-50.ec2.internal	□ IP-10-0-1-50.ec2.Internal	
AMIs	Answer private resource DNS name	Instance type	Elastic IP addresses
AMI Catalog	-	c3.large	34.197.120.45 [Public IP]

- d. The terraform script will pull the code from github repository https://github.com/sumitkumarsahoo-x21154589/airflow\_terrafoam.git and clone the data inside the EC2 webserver spin up above
- e. We can check it using ssh inside the EC2 server using git bash ssh by goin to the folder having test\_ec2\_key\_pair.cer file

	> x21154589-artifact-data-airflow_terrat	oam			
	Name	Status	Date modified	Туре	Size
<sup>o</sup> ersc	📒 .git	$\odot$	15-12-2022 05:31	File folder	
	ags 🔁 dags	$\odot$	15-12-2022 05:31	File folder	
*	scripts	$\odot$	15-12-2022 05:31	File folder	
*	source_data_for_research	$\odot$	15-12-2022 05:32	File folder	
*	🚞 temp	$\odot$	15-12-2022 05:31	File folder	
*	📒 terraform	$\odot$	15-12-2022 05:31	File folder	
*	📔 .gitignore	$\odot$	14-12-2022 20:10	GITIGNORE File	1 KB
*	.Rhistory	$\odot$	14-12-2022 20:10	RHISTORY File	0 KB
*	📔 docker-compose	$\odot$	14-12-2022 20:10	YML File	7 KB
li≉	Dockerfile	$\odot$	14-12-2022 20:10	File	1 KB
fɛ≯	InstructionToReviwer	$\odot$	15-12-2022 05:36	TXT File	1 KB
*	📔 README	$\odot$	14-12-2022 20:10	MD File	1 KB
	requirements	$\odot$	14-12-2022 20:10	TXT File	1 KB
	test_ec2_key_pair	$\odot$	13-12-2022 01:11	Security Certificate	2 KB

f. The ssh can done using following command:
>> ssh -i test\_ec2\_key\_pair.cer ubuntu@3.214.183.171

g. We can verify entire code is present or not in the EC2 using the command and make sure the root access is given to the folder.

>> cd / .

 Now the Apache Airflow web server is up and running on the EC2 instance it can be checked using the link <u>http://34.197.120.45:8080/home</u>
 Which is <u>http://<publicipaddressofec2>/home</u>

→ C ▲ Not see	cure   34.197.1 Anaconda   Indiv	20.45:8080/home	Airflow 🦲 Cond	ia-Issues 🧧 Logging 🗧 Cert	ifications 📃 Spark ETL-upsert	📕 Morningstar 📃 MyUs	년 🏚 🙂 😇 e 🖸 Java-Knowledge 📒 Java-Knowledge	😺 💓 🛸	•
Airflow DA	Gs Securi	ty – Browse –	Admin - Doo	25 v				07:56 UTC -	<b>AA</b> -
DAG Import Errors	5 (1)								>
AGs									
All 3 Active 2	Paused 1			Filter DAGs by tag			Search DAGs		
DAG	Owner	Runs 🕕	Schedule	Last Run 🕕	Next Run 🕕	Recent Tasks 🕕		Actions	Links
airflow_emr	airflow		@once	2022-12-12, 14:53:00 🕕				► Ō	
D airflow_emr_2	airflow	2 0	@once	2022-12-14, 18:48:52 🕕		00000		ÞŌ	
tutorial	airflow		1 day, 0:00:00	2015-06-10, 00:00:00 🕧	2015-06-11, 00:00:00 🕕	00000	000000000	ÞŌ	
c 1 > >>								Showing 1-3	3 of 3 DAG
/ersion: v2.2.4 Git Version: .release:2.2	.4+ee9049c056	36b2539a247687de05	if9cffa008f871						

This completes the deployment steps of EC2 and S3

## 5 Setup AWS Services

## 5.1 AWS redshift



1. Search for Redshift & Create Cluster



2. Select Production, cluster size dc2.large and Number of Nodes 1

aws Services Q Search	[Alt+S]	Ъ 🗘 🔍 м. v
Amazon Redshift ×	Cluster configuration	
Redshift serverless New	Cluster identifier This is the unique key that identifies a cluster.	
	redshift-cluster-1	
Provisioned clusters dashboard	The identifier must be from 1-63 characters. Valid characters are a-z (lowercase only) and - (hyphen).	
▼ Clusters	What are you planning to use this cluster for?	
Reserved nodes Snapshots	Production     Configure for fast and consistent performance at the best     price.     Origure for fast and consistent performance at the best     price.     Origure for learning about Amazon Redshift. This     configuration is free for a limited time if your     organization has never created an Amazon Redshift     cluster.	
Query editor		
Query editor v2 🛂	Choose the size of the cluster	
Queries and loads	• I'll choose	
	O Help me choose	
Datashares	Node type Info Choose a node type that meets your CPU, RAM, storage capacity, and drive type requirements.	
Configurations	Number of nodes Enter the number of nodes that you need.	
AWS Partner Integration		

3. Click on Create Cluster with other deafult config. This creates Cluster

aws Services Q Search		[Alt+S]	۵ 4	N. Virginia 🔻 x21154589-sumit-thesis
Amazon Redshift ×	General information			C
Provisioned clusters dashboard	Cluster identifier x21154589-sample-cluster Cluster namespace	Status Available Date created	Node type dc2.large Number of nodes	Endpoint T x21154589-sample-cluster.cuexqajw
▶ Clusters	95e7e918-e7f8-4b76-b280-75c558f85a02 Cluster configuration	December 13, 2022, 21:48 (UTC+00:00) Storage used	1	JDBC URL  JDBC URL  JDBC URL  ODBC URL  ODBC URL
Query editor Query editor v2 🛃 Queries and loads	Production	Multi-AZ No		ODSC URL
Datashares	Cluster performance Query monitori	ng Schedules Maintenance	Properties	
Configurations	<ul> <li>Recommendations (0) To improve performance and decrease operating</li> </ul>	g costs, the Amazon Redshift Advisor provides recom	mendations.	
<ul> <li>AWS Partner Integration</li> <li>Informatica Data Loader</li> </ul>	► Alarms (0)		► Events (1)	

4. Reate the Database table as per the out file schema

😑   📀	🧶   G   G   🧮   🐱   🔿   🕹	6   🛛   🏓   🛇   🛇	🐰   🔶   🤄   🐹   G	🔶   C   C   C   G   S	🔍 🔍 🔍 🖉 🖉 🖉	+ ~ - 0 ×
$\leftarrow \rightarrow$	C 🔒 us-east-1.console.aws.amazon.co	m/sqlworkbench/home?region=u	ıs-east-1#/client		☞ ☆	😊 😇 🔰 🗰 🗯 🗖 🚳 E
a Hands-C	On Python f 😑 Anaconda   Individu 📒 F	Python 📙 Airflow 🦲 Conda-Iss	ues 📙 Logging 📙 Certification	ns 📃 Spark ETL-upsert 📃 Mor	ningstar 📒 MyUse 📧 Java-Knowledge 📒 Java-	Knowledge Imported »
aws	Services Q Search		[Alt+S]		D 4°0 N	. Virginia 🔻 x21154589-sumit-thesis 🔻
) Database	Redshift query editor v2	+ E Untitled 1 ×	100 💽 Explain 💽 Isolat	ed session () x21154589-sa	ample-cl • x21154589_sample_cl •	8
Queries	Q Filter resources ♀ ✓ 중 x21154589-sample-cl① :	1 select * from tes	t_table			
Notebooks	<ul> <li>&gt; adev</li> <li>&gt; asample_data_dev</li> <li>&gt; sample_tlata_dev</li> <li>&gt; sample_cl</li> <li>&gt; public</li> <li>&gt; Tables</li> <li>4</li> </ul>					
() History	≣ stack_over_fl ♀ ×	E Result 1 (3)				1 Export - Chart x
	Field	🗆 fruit_id	name	color		
	A responseid	□ 1 □ 2	a b	d e		
	A mainbranch	3				
C	A employment					
۵	A country					Elapsed time: 207 ms Total rows: 3
Feedback	Looking for language selection? Find it in the	new Unified Settings 🔀			© 2022, Amazon Web Services, Inc. or its affiliates.	Privacy Terms Cookie preferences
-3°C Cloud	dv	Q Search 🔲 🚨	🝼 들 🔮 📦 🕸	🛛 🔼 🤹 💕 💶	😰 刘 🗣 🍇 🖬 🧄	∧ ENG (\$ \$ \$ \$ \$ \$ \$ \$ \$ \$ \$ \$ \$ \$ \$ \$ \$ \$ \$

# 6 Running Apache Airflow DAG

- In Visual Code Studio, after development changes are done, they can be pushed to GitHub using the command in the terminal: git remote add origin https://github.com/sumitkumarsahoo-x21154589/s.git git branch -M main git push -u origin main
- 2. Prerequisite of running DAG,
  - a. setting up source in S3.

b. Upload the <u>survey\_results\_public.csv</u> from Public Secondary Dataset (https://insights.stackoverflow.com/survey? ga=2.91719770.170077947.1671050547-2078385242.1671050547 )source file in x21154589\_input folder of S3 bucket

Buckets	x21154589_input/	Copy S3 U
Access Points		
Object Lambda Access Points	Objects Properties	
Multi-Region Access Points		
Batch Operations		
Access analyzer for S3	UDJECTS (2) Objects as the fundamental antitiar steroid in Amazon CE. You can use Amazon CE investory CP to get a list of all objects in your buckets. For others to access your objects you'll need to explicit	icitly grant them
	Outpets are the formalise the formation so. For an use Analosi so interiory may be a not on indepets any your bucket. For outpets to access your objects, you in need to expendence of the permissions. Learn more 🖉	icity grant area
Block Public Access settings for this account	or C C Copy 53 URI C Copy URL C Download Open C Delete Actions ▼ Create folder C Uplo	ad
Block Public Access settings for this account Storage Lens	or C Copy 53 URI C Copy URL C Download Open C Delete Actions ▼ Create folder C Uplo Q Find objects by prefix <	ad
Block Public Access settings for this account Storage Lens Dashboards	or Copy S3 URI Copy URL Download Open Z Delete Actions ▼ Create folder R Uplo Q Find objects by prefix Name ▲ Type ▼ Last modified ▼ Size ▼ Storage	ad 1 > {
Block Public Access settings for this account Storage Lens Dashboards AWS Organizations settings	or Copy S3 URI Copy URL Download Open Z Delete Actions ▼ Create folder R Uplo C Find objects by prefix Name Type ▼ Last modified ▼ Size ▼ Storage December 15, 2022, 08:01:17 (UTC+00.00) 0 B Standar	ad 1 > { e class

## c. Goto <a href="http://34.197.120.45:8080/home">http://34.197.120.45:8080/home</a> and Click on Admin Drop Down-> Connections

●   ③   ●   G   G   Ⅲ   ঊ   〇   ③	G   <b>Q</b>   <b>Q</b>   <b>Q</b>   <b>Q</b>   <b>X</b>   <b>♦</b>   <b></b> €	M   G   �   Œ   Œ   G   δ   ♥   ●   A   O	0  x   +
← → C ▲ Not secure   34.197.120.45:8080	I/home		🖻 🛧 😃 🖲 🔰 🗰 🗯
3. Hands-On Python f 🧿 Anaconda   Individu 📒 I	Python 🧧 Airflow 🧧 Conda-Issues 📒 Logging 📒	Certifications 📒 Spark ETL-upsert 📒 Morningstar 📒 MyUse 💶 Java	-Knowledge 📒 Java-Knowledge 🧧 Imported 🛛 »
Airflow DAGs Security Br	rowse - Admin - Docs -		08:31 UTC - AA -
DAG Import Errors (1)	Variables Configurations		>
DAGs	Connections Plugins Providers		
All 3 Active 2 Paused 1	Pools In DAGs by tag XComs		Search DAGs
DAG Owner Runs	Schedule Last Run	Next Run 🌒 Recent Tasks 🕥	Actions Links
emr airflow_emr	2022-12-12, 14:53:00	• 0000000	
emr airflow_emr_2 airflow	2022-12-14, 18:48:52	• • • • • • • • • • • • • • • • • • • •	
tutorial airflow	1 day, 0:00:00 2015-06-10, 00:00:00	2015-06-11, 00:00:00	
« ( <mark>1</mark> ) »			Showing 1-3 of 3 DAGs
Version: v2.2.4 34.197.120.45:8080/connection/list/ ee9049c0566b2539a2	47687de05f9cffa008f871		
-3°C Cloudy	Q Search 🔲 🖸 🤡 😂	😆 💶 🖪 🤹 🥵 🖷 🗟 🍕 🦫 🌸	

→ C ▲ Not sec							
	ure   34.197.120.45	:8080/connectio	n/list/			ie 🖈 🐵	U 😾 🗰 🏞 🔲 (
lands-On Python f 📀	Anaconda   Individu	Python	Airflow 🧧 Conda-Issues	📙 Logging 📙 Certifications 📃 Spark ETL-upsert 📃 Morningstar 📒 MyUse	Java-Knowledg	e 📙 Java-Knowled	ge 🔜 Imported
Airflow DAG	Gs Security -	Browse	Admin - Docs -				08:32 UTC - AA
t Connection							
Search -							
+ Actions - +							Record Count: 3
	Conn Id 🛿	Conn Type 1	Description 1	Host 1	Port 1	Is Encrypted	Is Extra Encrypted 1
	aws_default	aws				False	False
	emr_default	emr				False	False
	redshift_default	redshift	config to access reshift	x21154589-sample-cluster.cuexqajwmseq.us-east-1.redshift.amazonaws.com	n 5439	False	False

Version: v2.2.4 Git Version: .release:2.2.4+ee9049c0566b25	539a2476	87de05f9cffa0	08f871																			
Watchlist -1.59%		Q, Search		D	<b>@</b>	-	6	۲	w	<b>.</b>	49	2	×	ŋ.	-	<b>D</b>	0	^	ENG IN	\$ 4) 20	08 15-12-20	32 22 10

d. Setup connection string for AWS login , EMR spin up and close(for cost saving) and redshift connection to move proceesed data from S3 bucket <u>emrbuckettestsumit</u>/ 21154589\_output/

nection Type *	Amazon Web Services * Connection Type missing? Make sure you've installed the corresponding Airflow Provider Package.
scription	
ıt	
ema	
jin	
sword	
t	
ra	(?aws_access_key_id": "AKIAXLJZTDDBNRIGCZXUM", "aws_secret_access_key", "DSZbgTSC/IAAOInku3UbsWk09CVRJwWWULKI3Xm", "region_name", "us-east-1")
ive 🖹 Test 🖌 🗧	

Airflow DAGs Security	- Browse - Admin - Docs -	08:35 UTC -	AA
lit Connection			
connection Id *	em_oesut		
connection Type *	Amazon Elastic: MapReduce • Connection Type missing? Make sure you've installed the corresponding Akrlow Provider Package.		
escription			,
ost			
chema			
ogin			
issword			
ərt			
tra	("Name": "terv_lesting_airlow", "LogUri" "33/IA21154599embucket/", "Instances" ("EcXKeyName": "terv_lest_airlow_ec2", "Ec2Subnetid": "subnet-0ddact6536646678bc", "MasterinstanceType": "Inst.Atarge",	", "SlaveInstanceType": ServiceRole":	,
ave 🖹 🛛 Test 📌 🗧 🗧			
INFY News Bearish	📑 (9. Seach) 🖬 🗊 💇 📮 🍪 👾 🖏 📮 🗳 🍄 🚏 🖉 🏂 🏂 🕎 🔺	ENG IN 중 데) 絤	08 15-12-20
INFY News Bearish	■       Q. Search       ■       ■       ●       ●       ■       ●	ENG (\$ 0) (2) IN (- T) (2) (2) (2) (2) (2) (2) (2) (2) (2) (2) (2) (2) (2) (2)	08 15-12-20
INFY News Bearish → C G G		ENG (Construction) IN (Construction)	08
INFY News Bearish ⇒ C ▲ Not secure   3 lands-On Python f_ ▲ Ancone Airflow DAGs Security Micronection		ENG IN © 40 🏍 U U U H dge Imported 08:35 UTC -	08 15-12-20
NPY News Bearish → C ▲ Not secure 3 Hands-On Python L ④ Anaconc Airflow DAGS Security dit Connection	Search  Searc	ENG IN © (1) (1) O	08 15-12-20
NFY News Bearish	Image: Constraint of the second se	ENG IN $(* 4)$ (* (* - 1) (* - 1)	08 15-12-2( 0
INFY News Bearish	Image: Search       Image: Search<	ENG IN © (1) (200	08 15-12-20
INFY News Bearish	Image: Search       Image: Image	ENG IN © (1) (200	088 15-12-24
INFY         News Bearish         Image: Construction of the secure of t	Image: Construction Vedity         Image: Constructin Vedity         Ima	ENG IN $(a)$ (b) (b) (a) (c)	08 15-12-20 •••••••••••••••••••••••••••••••••••
NPY News Bearish C C C	Search	ENG IN © (1) (200	088 15-12-20 0
NEY News Bearish	Search     Search <td>ENG IN © (1) (1) © (1) (1) 00:35 UTC -</td> <td>083</td>	ENG IN © (1) (1) © (1) (1) 00:35 UTC -	083
NFY     News Bearish     C C C	Search S	ENG IN © d) () T T T T T T T T T T T T T T T T T T T	088
NPY Heven Bearish → C ▲ Not secure 3 Hands-On Python f_ ▲ Anaconc ▲ Arflow DAGs Security S	Search  Searc	ENG IN © (1) (months) 08:35 UTC -	088
NFY News Bearish G G G Ⅲ C ▲ Not secure 3 Hands-On Python L ▲ Anaconc Airflow DAG Security dt Connection 1 Connection 1 Connection 7ype * Connection 7ype * Connec	Search  Searc	ENG IN © (1) (1) T T T T T T T T T T T T T T T T T T T	080

- 3. Now we are set to run dag
- 4. Running the DAG:
  - a. Broef description of what the DAG will do:
    - i. First it will pull data from S3 bucket <u>emrbuckettestsumit</u>/ 21154589\_input/ <u>survey\_results\_public.csv</u>
    - ii. Then it will spin up EMR cluster
    - iii. It will run the PySpark code for transformation on input dataset
    - iv. Then it will do checks if the first two steps are successful or not
    - v. It will store the processed data in the <u>emrbuckettestsumit</u>/ 21154589\_output/ folder in parquet format



vii. The final step is the parquet file will be moved to Redshift for data analysis Before running the DAG we need to move the source data into s3 bucket emrhuckettersumity 21154589, input:

b.

maron 67 V				
11a2011 55 X	Amazon S3 > Buckets > emrbuckettestsumit	t > x21154589_input/		
<mark>ickets</mark> cess Points	x21154589_input/			🗇 Copy S3 UI
:ti-Region Access Points ch Operations ess analyzer for S3	Objects (2) Objects are the fundamental entities stored in Amazo	n 53. You can use Amazon 53 inventory 🛃 to get a list of all objects in yo	ur bucket. For others to access your objects, you'll i	need to explicitly grant them
k Public Access settings for account	C Copy S3 URI Copy	URL Download Open Z Delete	Actions <b>v</b> Create folder	R Upload < 1 > €
rage Lens hboards	Name	▲ Type v Last modified	v Size v	Storage class
5 Organizations settings	public_data_survey.csv	csv December 15, 2022, 08:01:17 (UTC+	00:00) 0 B	Standard
	survey_results_public.csv	csv December 14, 2022, 13:24:50 (UTC+	00:00) 76.2 MB	Standard

c. Go to link : <u>http://34.197.120.45:8080/home</u> and run the DAG as highlighted in yellow below:

Airflow DA	Gs Secur	ity - Browse -	Admin - Do	CS -				08:24 UTC -	<b>AA</b> -
DAG Import Errors	s (1)								>
DAGs									
All 3 Active 2	Paused 1			Filter DAGs by tag			Search DAGs		
DAG	Owner	Runs 🕕	Schedule	Last Run 🛞	Next Run 🕕	Recent Tasks 🔘		Actions	Links
airflow_emr	airflow		@once	2022-12-12, 14:53:00 🕚		00000000		ÞŌ	
emr	airflow		@once	2022-12-14, 18:48:52 🕕		0000000000	000000	•	
tutorial	airflow	0	1 day, 0:00:00	2015-06-10, 00:00:00 🕚	2015-06-11, 00:00:00	0000000000			
« ( <b>1</b> ) »								Showing 1-3	of 3 DAGs

- Version: v2.2.4
- d. This would run the dag airflow\_emr\_2

e. We can check the live status of DAGs by click on **airflow\_emr\_2 and click on Graph view** 

Hands-Off Fytholi I	nda   Individu	Airflow Cond	Ja-Issues Subging	Certifications Spark ETL-up	isert Morningstar	MyUse 🖬 Jav	/a-Knowledge	ava-Knowledge	Imported	
Airflow DAGs Securit	ty- Browse- Admin-	Docs							08:41 UTC -	AA
DAG Import Errors (1)										3
AGs										
All Active 2 Paused 1			Filter DAGs by tag				Search DAGs			
DAG	Owner	Runs 🔘	Schedule	Last Run 🛞	Next Run 🔘	Recent Tasks 🔘			Actions	Lin
airflow_emr	airflow	000	gonce	2022-12-12, 14:53:00 🌒			00000	0000		
airflow_emr_2 t-in Spark app on Amazon EMR	airflow	000	Gonce	2022-12-14, 18:48:52 🍈			0000		•	
tutorial	airflow	000	1 day, 0:00:00	2015-06-10, 00:00:00 🍈	2015-06-11, 00:00:00 🔘		00000			
									Showing 1-3	3 of 3 D

Graph view

• • • • • • • • • • • • • • • • • • •	+	~ -	o ×
← → C 🕼 Not secure   34.197.120.45.8080/graph?dag_id=airflow_emr_2 Q 住 ☆	💡 💩 😇 ।	<b>V 🕷 🛪</b> 🗆	1 🚳 E
🚊 Hands-On Python f 🥥 Anaconda   Individu 🚦 Python 📋 Airflow 📋 Conda-Issues 🖺 Logging 🖺 Certifications 🗎 Spark ETL-upsert 📳 Morningstar 🛔 MyUze 🚥 Java-Knowledge 📒 Ja	ava-Knowledge	Imported	39
XAIrflow DAGs Security- Browse- Admin- Docs-		08:42 UTC -	<b>AA</b> -
DAG: airflow_emr_2 Run built-in Spark app on Amazon EMR	failed Schedule	e: @once Next R	un: None
👷 Tree 🥞 Graph 🔂 Calendar 😰 Task Duration 🛱 Task Tres 👱 Landing Times 📃 Ganit 🔥 Details ↔ Code			• •
2022-12-14T18.48.552         Runs         25         V         Run         manual_2022-12-14T18.48.52.355013+00.00         V         Layout         Left > Right         V         Update	Find Task		
EmildEspectantial EmiliationCyantal EmiliationCyantal Emiliational Statestantogenetal (structure) (or to restable) (or to res	pstream_failed skipped	scheduled deferred	] no_status
		Auto-refrest	C
create_job_flow $\rightarrow$ add_steps $\rightarrow$ watch_step $\rightarrow$ transfer_s3_to_redshift			
🗢 3°C 📲 🔍 Search 🖬 💭 💇 🐂 🍪 🖷 🗖 🤹 🕸 📲 🗟 🔌 🦞 🖳 🖉	∧ ENG IN	奈 Q1) 🍎 15-	08:42 12-2022 10

## We can check the staus and logs by clicking on the steps in rectangle shape and clicking on logs

🌻   🎯   🌻   G   G   🧮   🐻   🖓   🥸   G   🕸   🏕   🏟	D   D   D   D   D   O   M   O   M   O   C   C   C   C	\$ \$ \$ \$ \$ \$ \$ \$ \$ \$ \$ \$ \$ \$ \$ \$ \$ \$ \$
← → C ▲ Not secure   34.197.120.45:8080/graph?dag_id=airflow_emr_i	2	Q 🖻 🛧 🤩 😇 💆 🗮 🗯 🖬 🌒 :
🧸 Hands-On Python f 😑 Anaconda   Individu 📒 Python 📒 Airflow 📒 Conc	da-Issues 🧧 Logging 🧧 Certifications 📑 Spark ETL-upsert 📒 Morningstar	🧧 MyUse 💶 Java-Knowledge 📒 Java-Knowledge 📒 Imported 🛛 »
Airflow DAGs Security- Browse- Admin - Docs-	Task Instance: watch step ×	08:44 UTC - 🤼 -
• DAG: airflow_emr_2 Run built-in Sperk app on Amazon EMR	al: 2022-12-14, 18:48:52 UTC	failed Schedule: @once Next Run: None
👷 Tree 🌱 Graph 🔂 Calendar 😰 Task Duration 📑 Task Tries 📥 Landing	Instance Details Rendered Log All Instances Filter Upstream	• •
2022-12-14T18.48.59Z     Runs 25 v Run manual_2022-12-14T18.4	Download Log (by attempts): 1	Find Task
Emilialityscherated EmiliaationOperator Emilianderener S315-ResentCoperator	Task Actions	y mana lake w.tr.sty w.tr.stated water. take supped scientif attand so.states
	Ignore All Deps Ignore Task State Ignore Task Deps Run	Auto-refresh C
	Past Future Upstream Downstream Recursive Failed Clear	
	Past Future Upstream Downstream Mark Failed	
	Past Future Upstream Downstream Mark Success	
	Close	
34.197.120.45:8080/log?dag_id=airflow_emr_2&task_id=watch_step&texecution_date=2022-12-141	18%3A48%3A52.395013%2800%3A00	ENC 08:44
Cloudy Q Search	🗩 💽 🚔 🔄 🖌 🦏 🚮 👘 🦓	× ↓ × × × × × × × × × × × × × × × × × ×

Log View

	o ×
🗧 🔶 🕐 🔺 Not secure   34.197.120.455.0080/log?dag_id=airflow_ermr_28.task_id=watch_step8xexecution_date=2022-12-14T18%3A48%3A52.395013%2800%3A00 🔍 🛱 🛠 🙄 😈 👹 🔅	🗆 🚯 :
🤹 Hands-On Python f 🔕 Anaconda   Individu 📒 Python 📒 Airflow 📒 Conda-Issues 📒 Logging 📒 Certifications 📑 Spark ETL-upsert 📑 Morningstar 📑 MyUse 💶 Java-Knowledge 📑 Java-Knowledge 📑 Imported	30
Kairflow DAGs Security Browse Admin Docs 08:45 UTC	<b>AA</b> -
DAG: airflow_emr_2 Run built-in Spark app on Amazon EMR	dule: @once
🝨 Tree 📲 Graph 🗔 Calendar 😰 Task Duration 🛱 Task Tries 👱 Landing Times 🚍 Gantt 🛕 Details ↔ Code	• •
Task Instance: watch_step at 2022-12-14, 18:48:52	
Δ Task Instance Details ↔ Rendered Template 🗮 Log 🛱 XCom	
Log by attempts	
Jump To End	Toggle Wrap
*** Reading local file:/opt/airflow/logs/airflow_ger_2/watch_ttp/2022-12-14T18.48152.395031400.00/1.log [2022-12-14.118:48159 UTC] (taskinstance.yi037) NFO - Dependencies all met for (Taskinstance: airflow_ger_2.watch_step manual_2022-12-14T18:48152.395013+00:00 [qurued]) [2022-12-14.118:48159 UTC] (taskinstance.yi124) NFO - Dependencies all met for (Taskinstance: airflow_ger_2.watch_step manual_2022-12-14T18:48152.395013+00:00 [qurued]) [2022-12-14.118:48159 UTC] (taskinstance.yi124) NFO - Dependencies all met for (Taskinstance: airflow_ger_2.watch_step manual_2022-12-14T18:48152.395013+00:00 [qurued]) [2022-12-14.118:48159 UTC] (taskinstance.yi124) NFO - Dependencies all met for (Taskinstance: airflow_ger_2.watch_step manual_2022-12-14T18:48152.395013+00:00 [qurued]) [2022-12-14.118:48159 UTC] (taskinstance.yi124) NFO - Dependencies all met for 1 [2022-12-14.118:48159 UTC] (taskinstance.yi1240) NFO - Dependencies all met for 1	
[2022-12-14, 11:04)9 UTC] [tskinstee.py:1244] 1MO - Executing (Tesk[Exercise]seesor): watch_step on 2022-12-14 11:04:05.2.39811400.000           [2022-12-14, 11:04)9 UTC] [tskinstee.py:1244] 1MO - Executing (Tesk[Exercise]seesor): watch_step on 2022-12-14 11:04:05.2.39811400.000           [2022-12-14, 11:04)9 UTC] [tskinstee.py:1244] 1MO - Executing (Tesk[Exercise]seesor): watch_step on 2022-12-14 11:04:05.2.39811400.000           [2022-12-14, 11:04:05 UTC] [tskinstee.py:1244] 1MO - Executing (Tesk[Exercise]seesor): watch_step on 2022-12-1411:04:05.2.39811400.000           [2022-12-14, 11:04:05 UTC] [tskinstee.py:1244] 1MO - Executing (Tesk[Exercise]seesor): watch_step on 2022-12-1411:04:05.2.39811400.000           [2022-12-14, 11:04:05 UTC] [tskinstee.py:130] 1MO - Executing (Tesk[Exercise]seesor): watch_step namel_2022-12-1411:04:05.2.39811400.000           [2022-12-14, 11:04:05 UTC] [tskinstee.py:130] 1MO - Executing the following env vars:           Attria.(CT_DAG_DAGEs***           Attria.(CT_DAG_DAGEs****           Attria.(CT_DAG_DAGEs****           Attria.(CT_DAG_DAGEs****           Attria.(CT_DAG_DAGEs****           Attria.(CT_DAGEs****	', '/tmp/tmp4
Course	08:45 15-12-2022 1



# 7 Data Analysis in Redshift:

1. Connect to the Database that is create using Airflow connection string

Hands-On P	Python f 🙆 Anaconda   Individu 🧾	Python 🧧 Airflow	🧧 Conda-Issues 📒 Logging 📒 Certifications 📒 Spark ETL-upsert 📒 Morningstar 📒	MyUse 🚺 Java-Knowled	lge 🧧 Java-Knowledge	Imported
ws III	Services Q. Search		Edit connection for x21154589-sample-cluster	× D 4	🕐 N. Virginia 🔻	x21154589-sumit-the
atabase	<ul> <li>× 21154589-sample-cluster</li> <li>&gt; dev</li> <li>&gt; sample_data_dev</li> <li>&gt; x21154589 sample cluster</li> </ul>		Authentication Learn more  Federated user The principal tags of your IAM role or user must provide the connection details. You configure these tags in IAM or your identity provider (IdP).			
			Temporary credentials     The query editor 1/2 generates a temporary password to connect to the database.     Database user name and password     Provide a tatabase user and password for the database that you are connecting to. The query editor			
			vs stores your redentiatis in AWS Secrets Manager on your behalt.   WS Secrets Manager  Choose a secret with endentials that are associated with the duster or that you created in AWS Secrets Manager.			
	7 in views		Database			
	stack_over_tlow_survey		The database name must be 1-64 characters. Valid characters are lowercase alphanumeric characters.			
			User name			
			x21154589_sample_cluster_db_user			
			Password The password must be 8-64 characters.			
			Show password			
ste .			Cancel			

3. Data Analysis using SQL queris

2.



8 Total Cost of ownership for the Setup infrastructure for an Year with the Hardware Configuration using AWS pricing calculator:

#### 12/14/22, 7:22 AM

AWS Pricing Calculator

Successfully updated Amazon EMR estimate.

Contact your AWS representative: https://aws.amazon.com/contact-us/

 $\times$ 

Export date: 12/14/2022	Language: English
-------------------------	-------------------

Estimate title: My Estimate

Estimate summary	timate summary		
Upfront cost	Monthly cost	Total 12 months cost	
0.00 USD	1,735.64 USD	20,827.72 USD	
		Includes upfront cost	

#### **Detailed Estimate**

Name	Group	Region	Upfront cost	Monthly cost
Business support plan	No group applied	All regions	0.00 USD	157.79 USD
Description: AWS-Su Config summary: Su Engineers for unlimit	upportEstimate upports 24/7 phone ted contacts, with a	e, chat, and emai and a response ti	l access to Cloud S me of less than 1	Support hour.
Amazon EC2	No group applied	US East (N. Virginia)	0.00 USD	77.45 USD
Description: Webser Config summary: Op Instances), EBS Stora	ver-Airflow perating system (Li age amount (8 GB)	nux), Quantity (1	), Pricing strategy	(On-Demand
Amazon EMR	No group applied	US East (N. Virginia)	0.00 USD	44.10 USD
Description: EMR- st Config summary: Nu (100 %Utilized/Mon Utilization (2 Hours/ Utilization (2 Hours/ (GB) (8), Job runtime	artstoped by Airflo umber of master EM th) Number of core Day) Number of ta: Day) Number of vC : (120 minutes), Tot	w 4R nodes (1), EC2 EMR nodes (2), 1 sk EMR nodes (1) PUs per job run ( tal ephemeral sto	2 instance (m5.xla EC2 instance (m5. , EC2 instance (m! (2), Amount of me orage per job run (	rge), Utilization xlarge), 5.xlarge), emory per job ru (GB) (8)
Amazon Redshift	No group applied	US East (N. Virginia)	0.00 USD	186.00 USD
Description: Reshift Config summary: No %Utilized/Month), P	applied DW odes (1), Instance ty ricing strategy (On	Virginia) /pe (dc2.large), U Demand), Additio	Itilization (On-Der	nand only) (10 ge (145 GB), Da

https://calculator.aws/#/estimate

1/2



## 9 Conclusion

Using the above said tools , software and AWS services it is established that we can create an Open source ETL Framework for Big Data using AWS services and orchestrating it using Terraform with cost -effective Cloud Solution Architecture