

DR. ALVIN'S PUBLICATIONS

USING APACHE SPARK IN AWS AND DATABRICKS

DR. ALVIN ANG



1 | PAGE

COPYRIGHTED BY DR ALVIN ANG
WWW.ALVINANG.SG

CONTENTS

I.	<i>What are We Trying to Do?</i>	4
II.	<i>Sign Up with Databricks (14 days Free Trial Full Account)</i>	7
III.	<i>Creating the Cross Account IAM Role</i>	11
A.	Add Credential Configuration from Databricks	11
B.	Go to IAM in AWS	12
C.	Create Role in AWS	12
D.	Enter the DataBricks Account ID in AWS	13
E.	Ignore the Step 2: Add Permissions and Just Click Next	14
F.	Enter a Role Name and Description	15
G.	Create Inline Policy in AWS New Role	16
H.	Copy JSON Code	17
I.	Paste JSON Code	18
J.	Copy Out the Amazon Resource Name [ARN] from the New Role (from AWS) into DataBricks	21
K.	Create a Configuration Name	22
IV.	<i>Create S3 Bucket in AWS</i>	23
V.	<i>Databricks Storage Configuration</i>	25
A.	Storage Configuration Name and Bucket Name	25
B.	Bucket Policy	26
C.	Finalize Configurations in Databricks	29
VI.	<i>Create Workspace in Databricks</i>	30
A.	Setup Workspace	30
B.	Create A Cluster	33
C.	Configure Cluster	34
D.	Meanwhile, head over to AWS EC2	35
VII.	<i>Important Note: Difference Between Account Mode vs Workspace Mode</i>	37
A.	Account Mode vs Workspace Mode	37
1.	Workspace Mode	38
2.	Account Mode.....	39

B.	You MUST Move from Account Mode to Workspace Mode	40
VIII.	Create Tables in Databricks.....	42
IX.	Create Notebook.....	45
A.	Attach / Detach Cluster... ..	46
B.	Start Spark Session.....	47
C.	Test Some Code.....	48
X.	VERY IMPORTANT: SHUT DOWN YOUR CLUSTER!!!.....	49
XI.	Appendix I: Signing Up with Databrick Community Edition [Free Forever but we won't be using this option in this Manuscript]	52
	About Dr. Alvin Ang	55

I. WHAT ARE WE TRYING TO DO?

The screenshot shows the Databricks website's 'AWS Pricing' page. At the top, there is a dark banner with a yellow icon of a building and a dollar sign, and text: 'Extended Time Databricks SQL Price Promotion - Save 40%+', 'Take advantage of our 15-month promotion on Serverless SQL and the brand new SQL Pro', and 'Learn More →'. The sidebar on the left contains navigation links: Platform, Solutions, Learn, Customers, Partners, Company, Try Databricks (highlighted in red), Watch Demos, Contact Us, and Login. The main content area has the heading 'AWS Pricing' and a red-bordered box containing the text: 'Databricks is deeply integrated with AWS security and data services to manage all your AWS data on a simple, open lakehouse'. Below this text are two buttons: 'Try for free' (red) and 'Learn more' (dark blue). To the right of the text is a decorative graphic consisting of a grid of red dots connected by thin red lines, forming a series of overlapping triangles. Two horizontal grey lines with black dots at their ends are also present.

- Realistically, Spark needs to run on clusters to do parallel computing.
- Databricks is an easy way of managing the complex Spark architecture.
- We shall use Databricks as the Jupyter Notebook interface [frontend] to run the codes and manage the clusters, but use AWS to form those clusters and store our data [backend].
- So Databricks is just going to help synchronize AWS.
- This manuscript first teaches how to do the setup between AWS and Databricks.
- Once the setup is done, we proceed to using Databricks to do Data Analysis.

Sign Up with AWS

aws Contact Us Support English My Account Sign In to the Console

Products Solutions Pricing Documentation Learn Partner Network AWS Marketplace Customer Enablement Events Explore More

AWS Free Tier Overview Free Tier Categories How to Create an Account Featured Offers for Business FAQs Terms and Conditions

AWS Free Tier

Gain free, hands-on experience with the AWS platform, products, and services

Learn more about AWS Free Tier

Create a Free Account

FEATURED
Startups may be eligible for AWS credits
AWS Activate provides eligible startups with a host of resources, including free AWS credits to spend on AWS services, and AWS Support.
[Sign up for Activate Today »](#)

Types Of Offers

Explore more than 100 products and start building on AWS using the Free Tier. Three different types of free offers are available depending on the product used. Click icon below to explore our offers.



Explore Free Tier products with a new AWS account.

To learn more, visit aws.amazon.com/free.



Explore Free Tier products with a new AWS account.

To learn more, visit aws.amazon.com/free.



Sign up for AWS

Confirm you are you

Making sure you are secure -- it's what we do.

We sent an email with a verification code to datafrens@gmail.com. (not you?)

Enter it below to confirm your email.

Verification code

Verify

Resend code

⊘ This password is publicly known through a data set leaked from a third party. Please try a different password. [FAQs](#)

Create your password

✔ It's you! Your email address has been successfully verified. **✕**

Your password provides you with sign in access to AWS, so it's important we get it right.

Root user password

Confirm root user password

Continue (step 1 of 5)



Free Tier offers

All AWS accounts can explore 3 different types of free offers, depending on the product used.



Always free
Never expires



12 months free
Start from initial sign-up date



Trials
Start from service activation date

Sign up for AWS

Contact Information

How do you plan to use AWS?

- Business - for your work, school, or organization
- Personal - for your own projects

Who should we contact about this account?

Full Name

Dr Alvin Ang

Phone Number

+65 97990262

Country or Region



Sign up for AWS

Secure verification

i We will not charge you for usage below AWS Free Tier limits. We may temporarily hold up to \$1 USD (or an equivalent amount in local currency) as a pending transaction for 3-5 days to verify your identity.



Billing Information

Credit or debit card number



AWS accepts all major credit and debit cards. To learn more about payment options, review our [FAQ](#)

Expiration date

Month Year

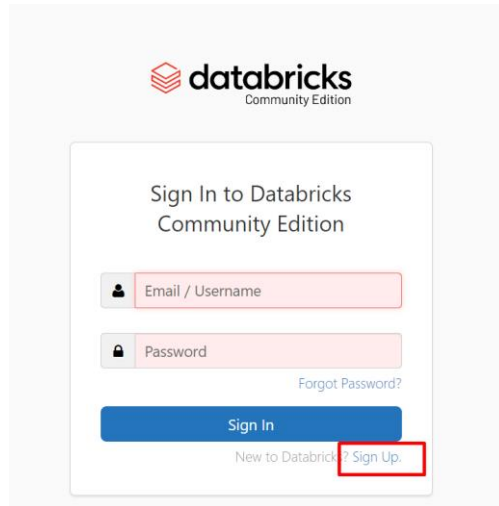
Cardholder's name

Billing address

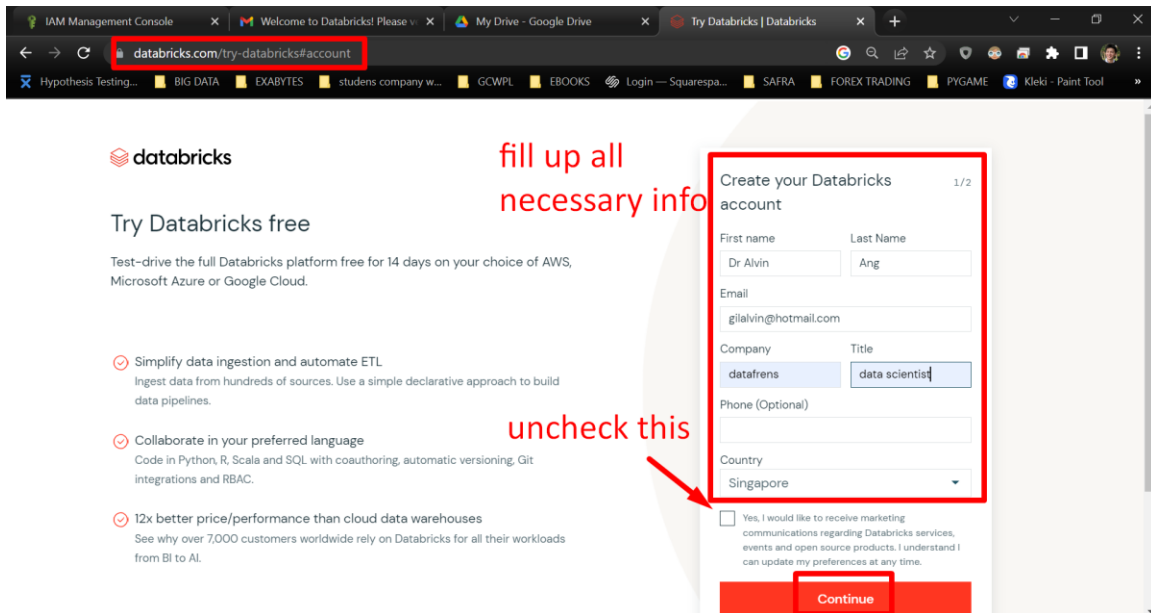
- Use my contact address
Blk 59 Tampines Central 7 #01-16
Singapore Singapore 528594
SG
- Use a new address

Verify and Continue (step 3 of 5)

II. SIGN UP WITH DATABRICKS (14 DAYS FREE TRIAL FULL ACCOUNT)



<https://www.databricks.com/try-databricks#account>



Let's choose AWS since we already have an account...but only free for 14 days TRIAL...

Try Databricks free

Test-drive the full Databricks platform free for 14 days on your choice of AWS, Microsoft Azure or Google Cloud.

- Simplify data ingestion and automate ETL
Ingest data from hundreds of sources. Use a simple declarative approach to build data pipelines.
- Collaborate in your preferred language
Code in Python, R, Scala and SQL with coauthoring, automatic versioning, Git integrations and RBAC.
- 12x better price/performance than cloud data warehouses
See why over 7,000 customers worldwide rely on Databricks for all their workloads from BI to AI.

Choose a cloud provider 2/2

- Amazon Web Services
- Microsoft Azure
- Google Cloud Platform

Continue

Don't have a cloud account?
Community Edition is a limited Databricks environment for personal use and training.

Get started with Community Edition

we won't choose this option here...

select this option if u want free forever but NO access to CLOUD (which means can't run SPARK in cluster)



Welcome to Databricks! Please verify your email address. [Inbox x](#)



Databricks <noreply@databricks.com>
to ALVINANG8888

1:33 PM (18 min)



Welcome to Databricks!

To complete your signup, please [verify your email address](#).

Or copy this link and paste it in your web browser:

<https://accounts.cloud.databricks.com/login?reset=password&username=ALVINANG8888%40GMAIL.COM&expiration=60000&token=10b15a1f0de2ba66343a1ada2d9883>

Select a subscription plan

Standard

Basic platform for your data analytics and ML workloads

- Databricks SQL
- Autoscaling
- Role based access control

Selected

Premium **Most popular**

Databricks SQL, cloud native security and autoscaling

- Databricks SQL
- Autoscaling
- Role based access control

Select Premium

Enterprise

Advanced compliance and security for mission critical data

- Databricks SQL
- Autoscaling
- Advanced compliance & security

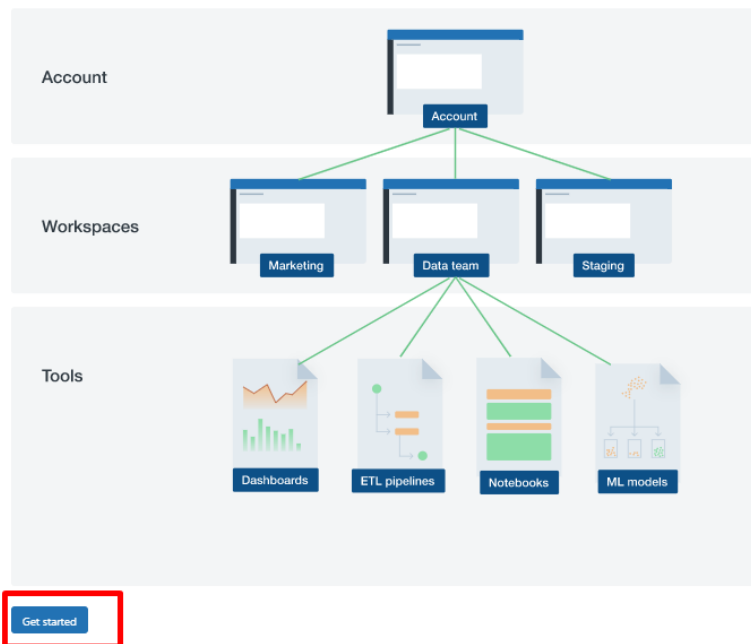
Select Enterprise

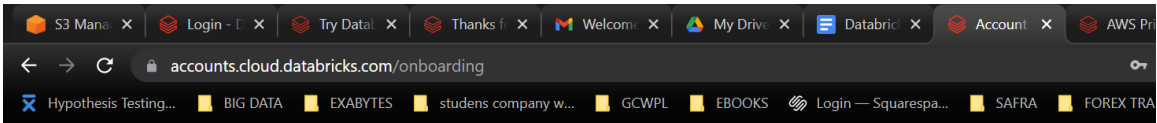
Your 14-day free trial starts when you click Continue. Thereafter, you will be charged at the list rates.

Continue



Workspaces are where teams solve the toughest data problems

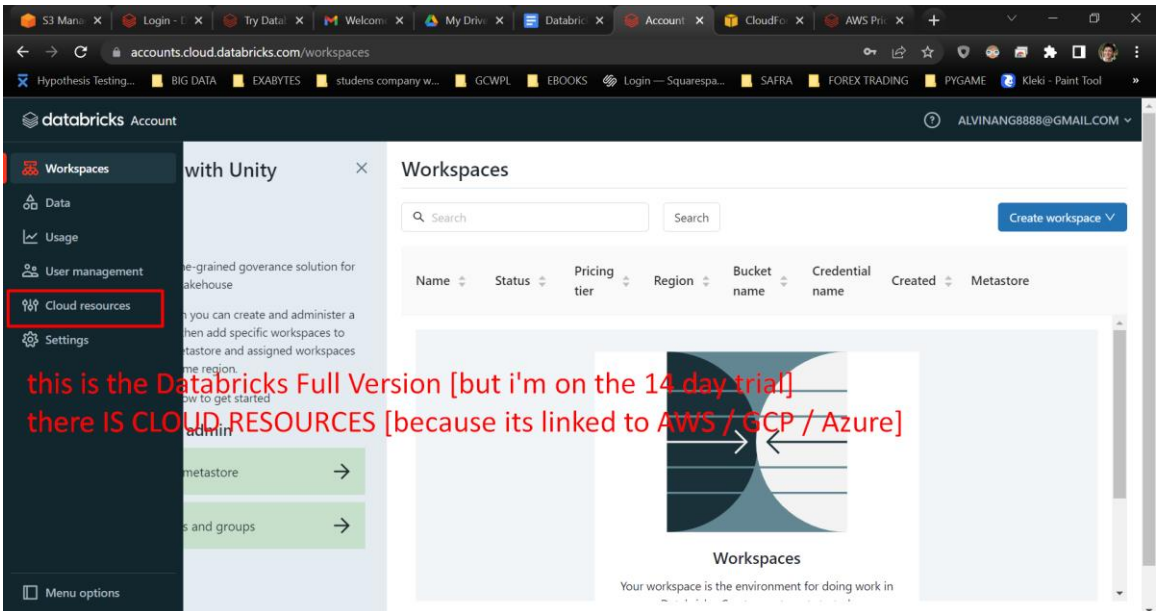




To proceed, please confirm you have the following

- ✓ An AWS account
- ✓ The password you used to setup your Databricks account
- ✓ A friendly name for your workspace

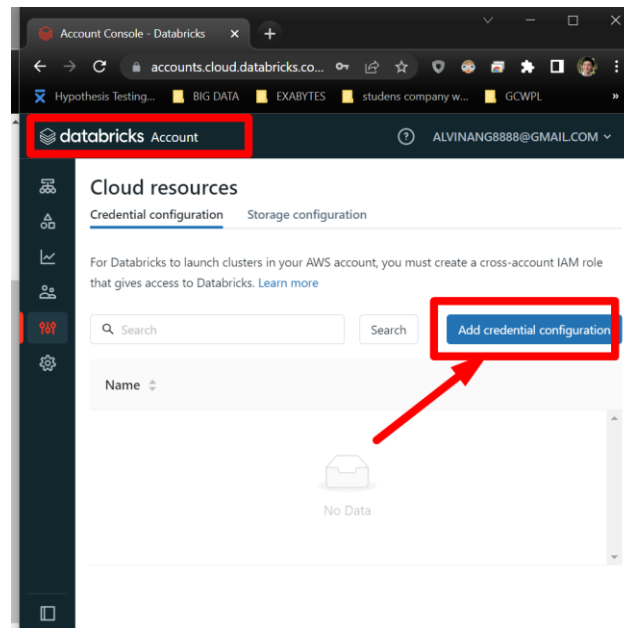
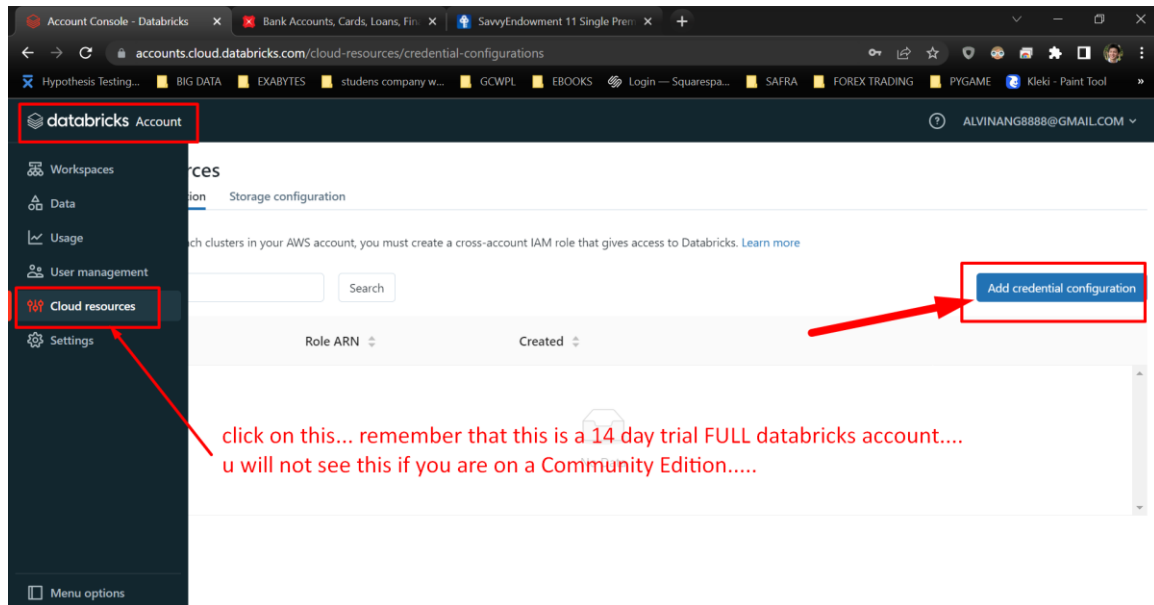
Confirm



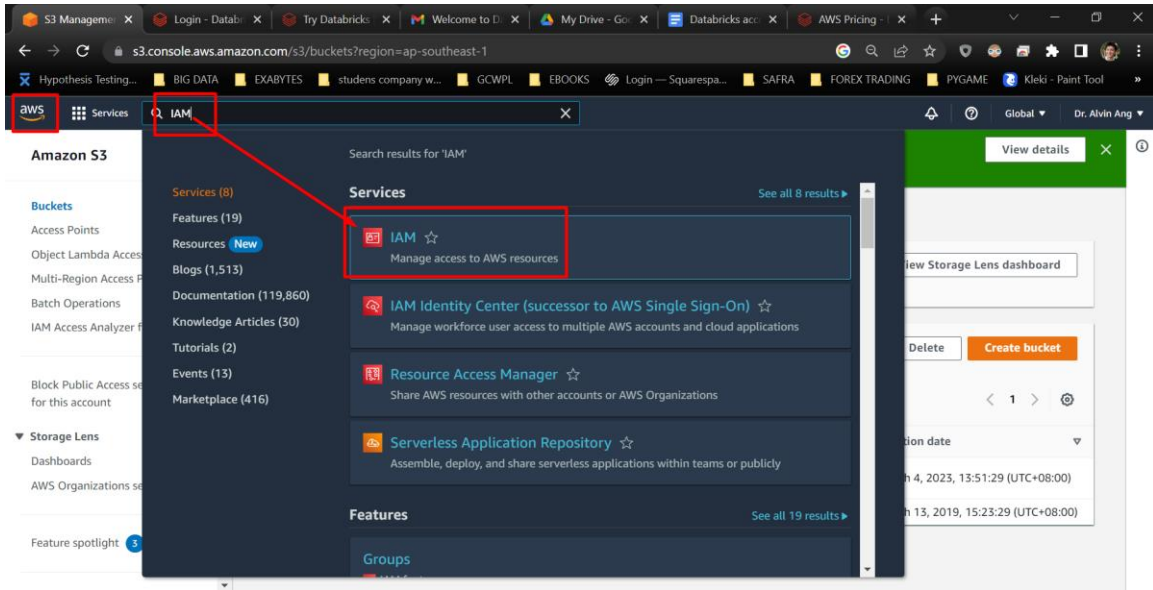
III. CREATING THE CROSS ACCOUNT IAM ROLE

<https://docs.databricks.com/administration-guide/account-api/iam-role.html#language-Databricks%C2%A0VPC>

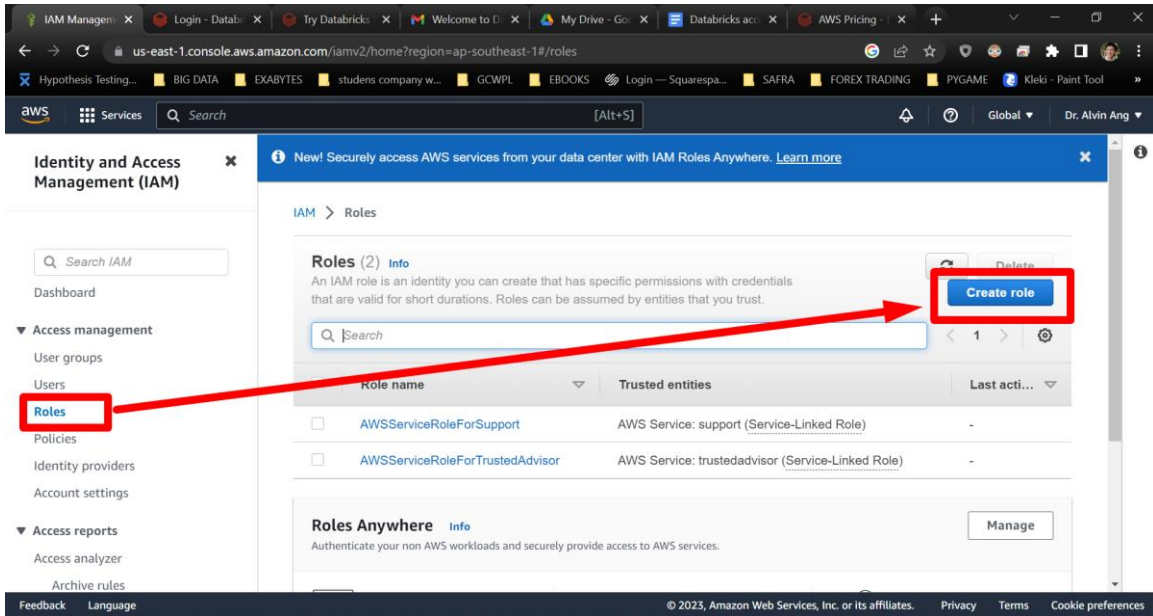
A. ADD CREDENTIAL CONFIGURATION FROM DATABRICKS



B. GO TO IAM IN AWS

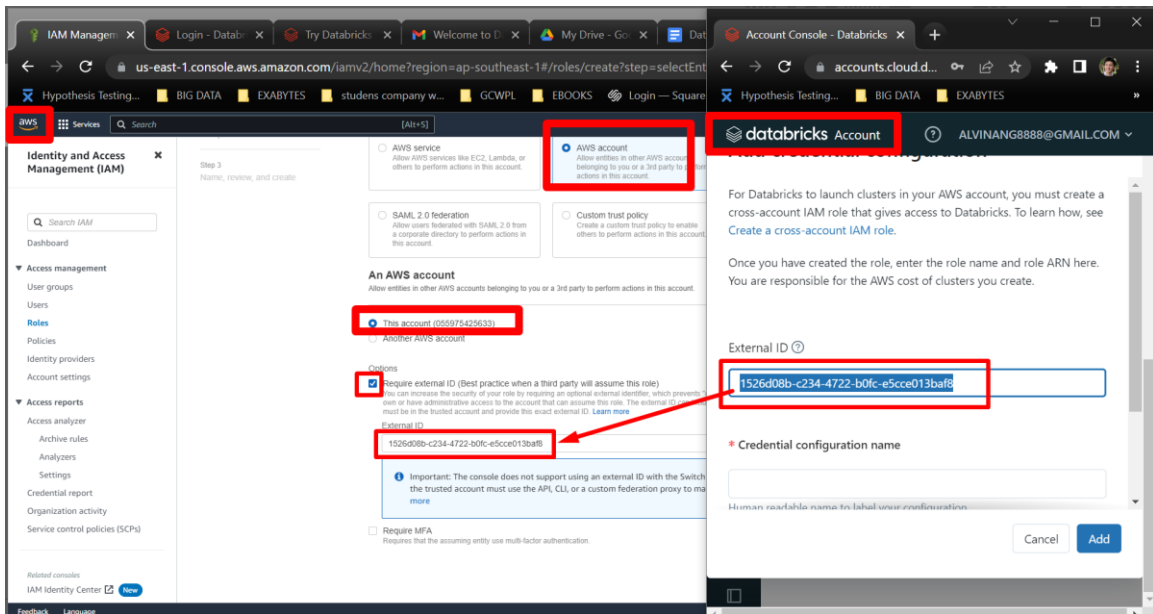
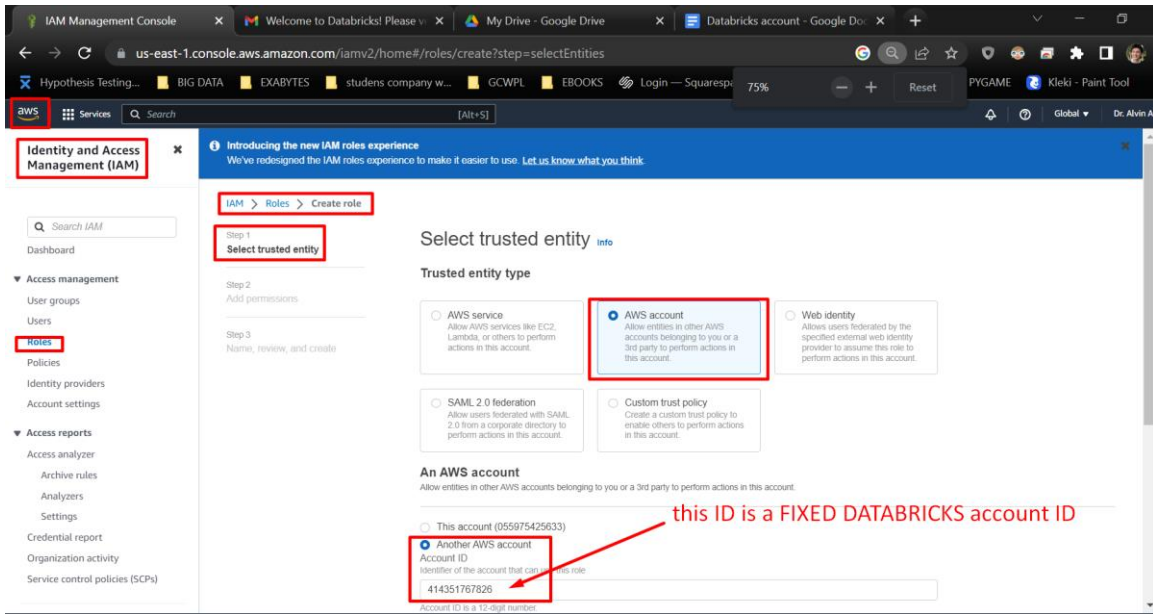


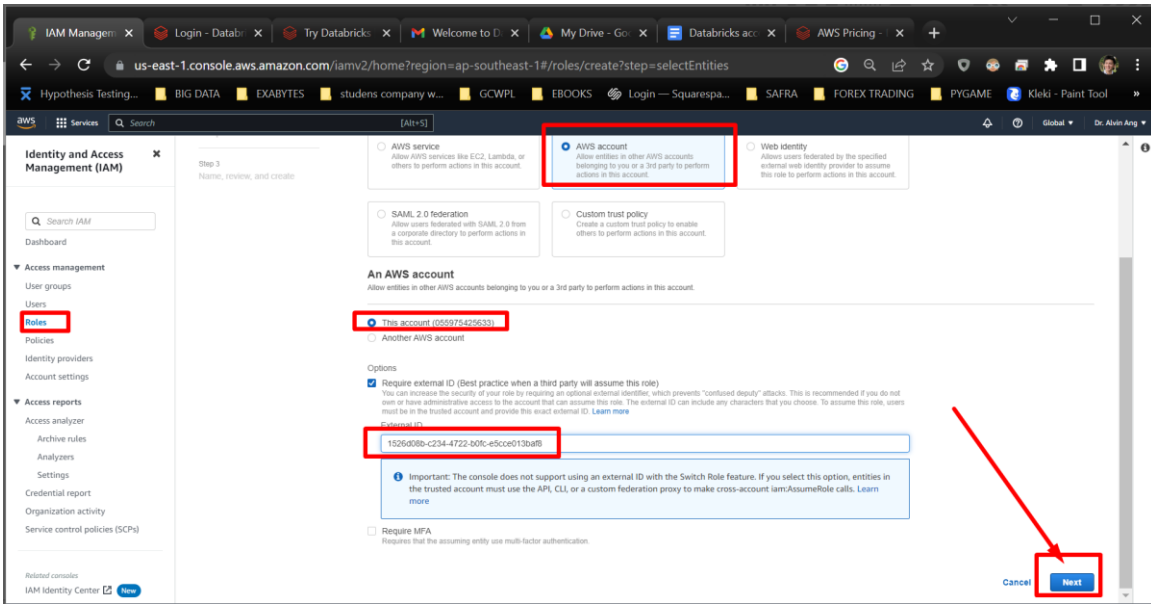
C. CREATE ROLE IN AWS



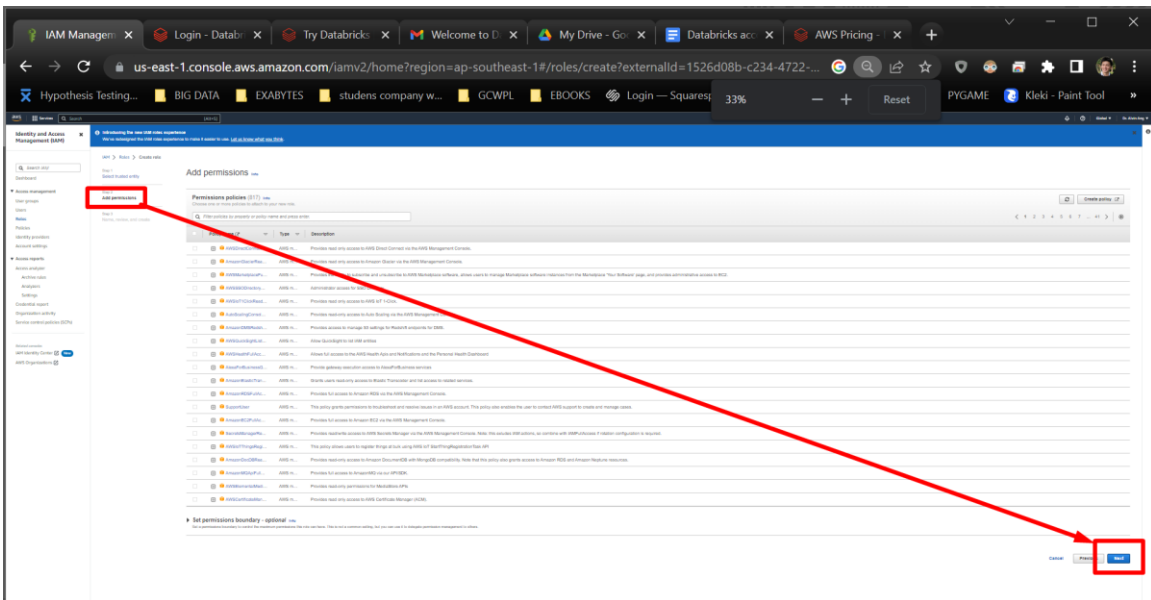
D. ENTER THE DATABRICKS ACCOUNT ID IN AWS

Databricks account ID = 414351767826





E. IGNORE THE STEP 2: ADD PERMISSIONS AND JUST CLICK NEXT



F. ENTER A ROLE NAME AND DESCRIPTION

The screenshot shows the AWS IAM console interface. On the left, a sidebar lists three steps: 'Step 1: Select trusted entity', 'Step 2: Add permissions', and 'Step 3: Name, review, and create'. The 'Step 3' label is highlighted with a red box. The main content area is titled 'Name, review, and create' and contains 'Role details'.

Role details

Role name
Enter a meaningful name to identify this role.

Maximum 64 characters. Use alphanumeric and '+=, @_-' characters.

Description
Add a short explanation for this role.

Maximum 1000 characters. Use alphanumeric and '+=, @_-' characters.

Step 1: Select trusted entities

The screenshot shows the 'Permissions policy summary' section of the AWS IAM console. A red arrow points from the text 'scroll down and let's create the ROLE' to the 'Create role' button. The 'Create role' button is highlighted with a red box.

scroll down
and let's create
the ROLE

Permissions policy summary

Policy name	Type	Attached as
No permissions added		

Tags

Add tags - optional [Info](#)

Tags are key-value pairs that you can add to AWS resources to help identify, organize, or search for resources.

No tags associated with the resource.

You can add up to 50 more tags.

G. CREATE INLINE POLICY IN AWS NEW ROLE

Identity and Access Management (IAM)

let's click on the new role we just created

Role name	Trusted entities	Last activity
AWSService_RoleForSupport	AWS Service: support (Service-Linked Role)	-
AWSService_RoleForTrustedAdvisor	AWS Service: trustedadvisor (Service-Linked Role)	-
databricks-aws-demo	Account: 414351767826	-

databricks-aws-role-demo

Permissions

Add permissions
Attach policies
Create inline policy

H. COPY JSON CODE

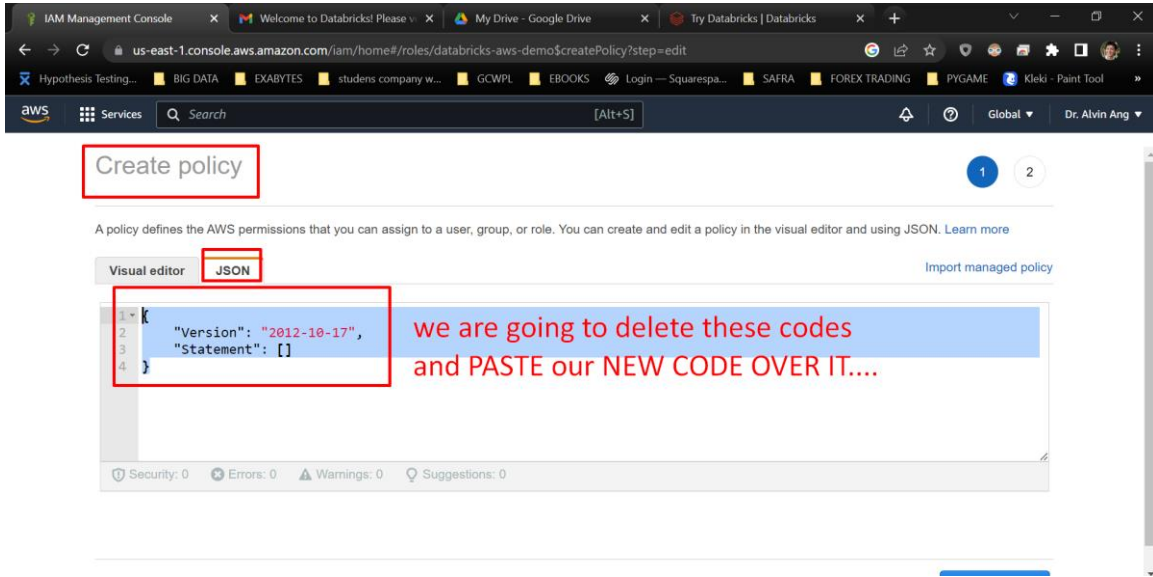
The screenshot shows a web browser displaying the Databricks documentation page. The URL in the address bar is `docs.databricks.com/administration-guide/account-api/iam-role.html#language-Databricks%C2%A0VPC`. The page content includes a search bar and a list of navigation links on the left. The main content area shows a section titled "c. Copy the access policy for deploying workspaces in a VPC that Databricks creates and configures in your AWS account." Below this, there is a code block containing JSON code. A red box highlights the "Copy" button next to the code. Red text annotations point to the "Copy" button and the JSON code, stating "JSON code can be found here (look for default policy)".

```
JSON
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Sid": "Stmt1403287045000",
      "Effect": "Allow",
      "Action": [
        "ec2:AllocateAddress",
        "ec2:AssociateDhcpOptions",
        "ec2:AssociateIamInstanceProfile",
        "ec2:AssociateRouteTable",
        "ec2:AttachInternetGateway",
        "ec2:AttachVolume",
        "ec2:AuthorizeSecurityGroupEgress",
        "ec2:AuthorizeSecurityGroupIngress",
        "ec2:CancelSpotInstanceRequests",
        "ec2:CreateDhcpOptions",

```

<https://docs.databricks.com/administration-guide/account-api/iam-role.html#language-Databricks%C2%A0VPC>

I. PASTE JSON CODE



```

{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Sid": "Stmnt1403287045000",
      "Effect": "Allow",
      "Action": [
        "ec2:AllocateAddress",
        "ec2:AssociateDhcpOptions",
        "ec2:AssociateIamInstanceProfile",
        "ec2:AssociateRouteTable",
        "ec2:AttachInternetGateway",
        "ec2:AttachVolume",
        "ec2:AuthorizeSecurityGroupEgress",
        "ec2:AuthorizeSecurityGroupIngress",
        "ec2:CancelSpotInstanceRequests",
        "ec2:CreateDhcpOptions",
        "ec2:CreateInternetGateway",
        "ec2:CreateNatGateway",
        "ec2:CreateRoute",
        "ec2:CreateRouteTable",
        "ec2:CreateSecurityGroup",
        "ec2:CreateSubnet",
        "ec2:CreateTags",
        "ec2:CreateVolume",
        "ec2:CreateVpc",
        "ec2:CreateVpcEndpoint",
        "ec2>DeleteDhcpOptions",
        "ec2>DeleteInternetGateway",
        "ec2>DeleteNatGateway",
        "ec2>DeleteRoute",
        "ec2>DeleteRouteTable",
        "ec2>DeleteSecurityGroup",
        "ec2>DeleteSubnet",
        "ec2>DeleteTags",
        "ec2>DeleteVolume",
        "ec2>DeleteVpc",
        "ec2>DeleteVpcEndpoints",
        "ec2:DescribeAvailabilityZones",
        "ec2:DescribeIamInstanceProfileAssociations",
        "ec2:DescribeInstanceStatus",
        "ec2:DescribeInstances",
        "ec2:DescribeInternetGateways",
        "ec2:DescribeNatGateways",
        "ec2:DescribePrefixLists",
        "ec2:DescribeReservedInstancesOfferings",
        "ec2:DescribeRouteTables",
        "ec2:DescribeSecurityGroups",
        "ec2:DescribeSpotInstanceRequests",
        "ec2:DescribeSpotPriceHistory",
        "ec2:DescribeSubnets",
        "ec2:DescribeVolumes",
        "ec2:DescribeVpcs",
        "ec2:DetachInternetGateway",
        "ec2:DisassociateIamInstanceProfile",
        "ec2:DisassociateRouteTable",
        "ec2:ModifyVpcAttribute",
        "ec2:ReleaseAddress",
        "ec2:ReplaceIamInstanceProfileAssociation",
        "ec2:RequestSpotInstances",
        "ec2:RevokeSecurityGroupEgress",
        "ec2:RevokeSecurityGroupIngress",
        "ec2:RunInstances",
        "ec2:TerminateInstances"
      ],
      "Resource": [
        "*"
      ]
    },
    {
      "Effect": "Allow",
      "Action": [
        "iam:CreateServiceLinkedRole",
        "iam:PutRolePolicy"
      ],
      "Resource": "arn:aws:iam:::role/aws-service-role/spot.amazonaws.com/AWSServiceRoleForEC2Spot",
      "Condition": {
        "StringLike": {
          "iam:AWSServiceName": "spot.amazonaws.com"
        }
      }
    }
  ]
}

```

IAM Management Console

us-east-1.console.aws.amazon.com/iam/home#/roles/databricks-aws-demo\$createPolicy?step=review

Create policy

1 2

Review policy

Before you create this policy, provide the required information and review this policy.

Name:

Maximum 128 characters. Use alphanumeric and "+, -, @, _" characters.

Summary

This policy defines some actions, resources, or conditions that do not provide permissions. To grant access, policies must have an action that has an applicable resource or condition. For details, choose **Show remaining**. [Learn more](#)

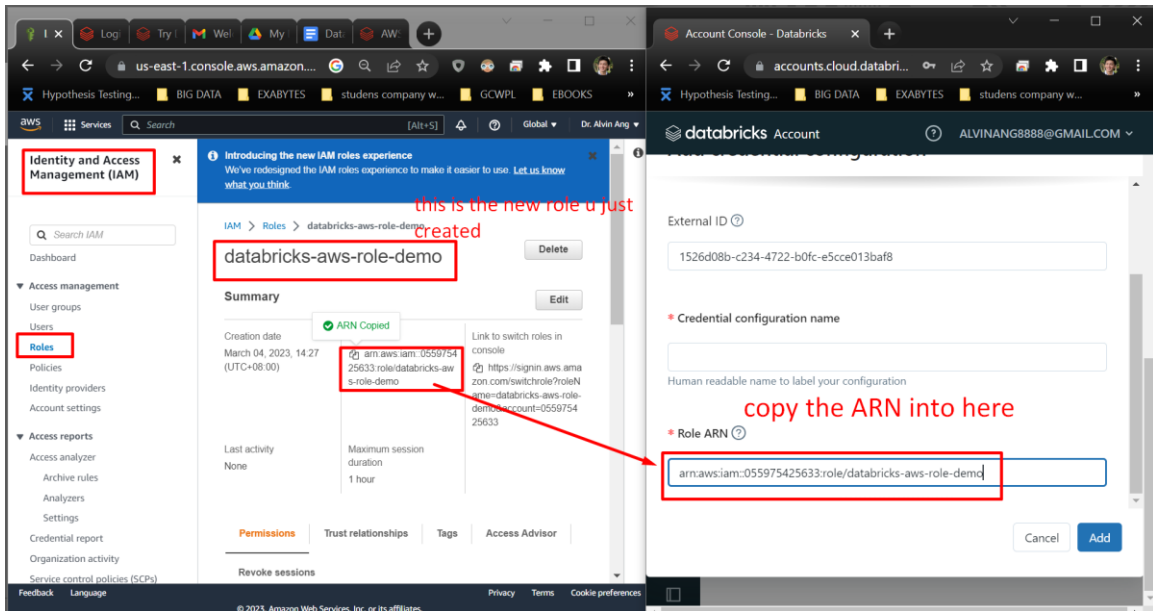
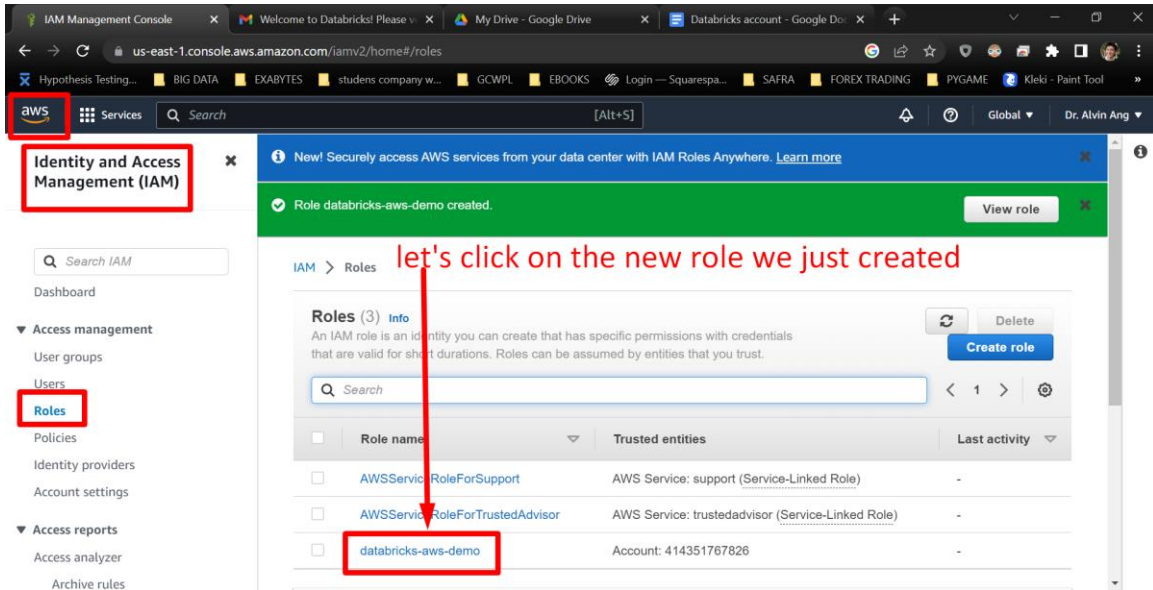
Filter

Service	Access level	Resource	Request condition
Allow (2 of 369 services) Show remaining 367			
EC2	Full, Tagging Limited	All resources	None
IAM	Limited	Path string like aws-service-roles/spot.amazonaws.com, RoleName string like AWSServiceRoleForEC2Spot	iam:AWSServiceName string like spot.amazonaws.com

* Required

Cancel Previous **Create policy**

J. COPY OUT THE AMAZON RESOURCE NAME [ARN] FROM THE NEW ROLE (FROM AWS) INTO DATABRICKS



K. CREATE A CONFIGURATION NAME

databricks Account

Add credential configuration

External ID ⓘ
1526d08b-c234-4722-b0fc-e5cce013baf8

* Credential configuration name
configuration-credentials-db-aws
Human readable name to label your configuration.

* Role ARN ⓘ
arn:aws:iam::055975425633:role/databricks-aws-demo

Cancel Add

Account Console - Databricks | Bank Accounts, Cards, Loans, Fin... | SavvyEndowment 11 Single Pre... | +

accounts.cloud.databricks.com/cloud-resources/credential-configurations

databricks Account now we shall configure this ALVINANG8888@GMAIL.COM

Cloud resources

Credential configuration Storage configuration

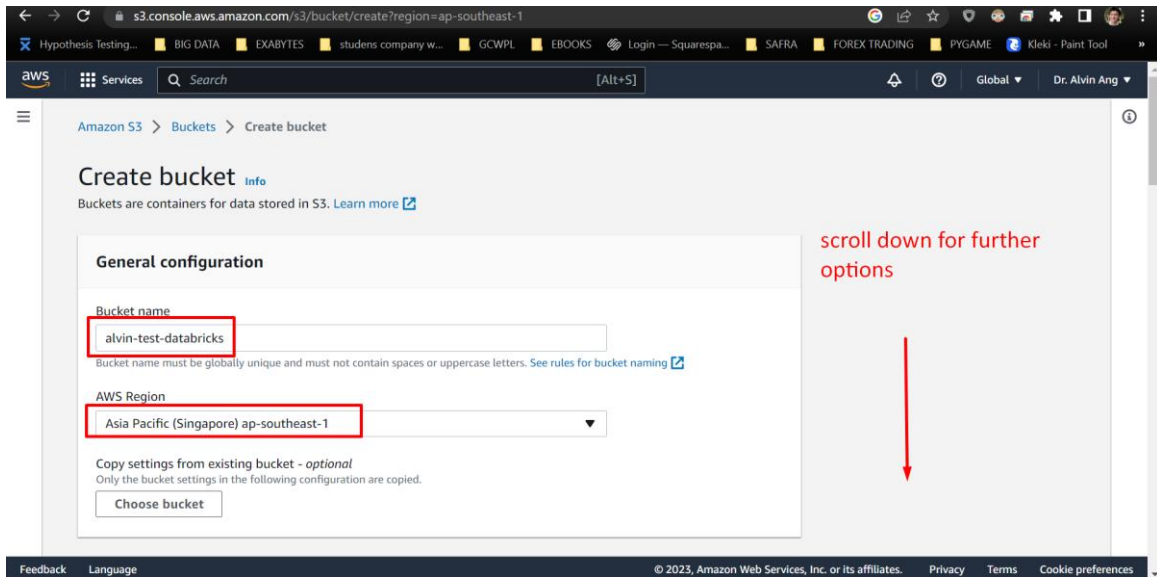
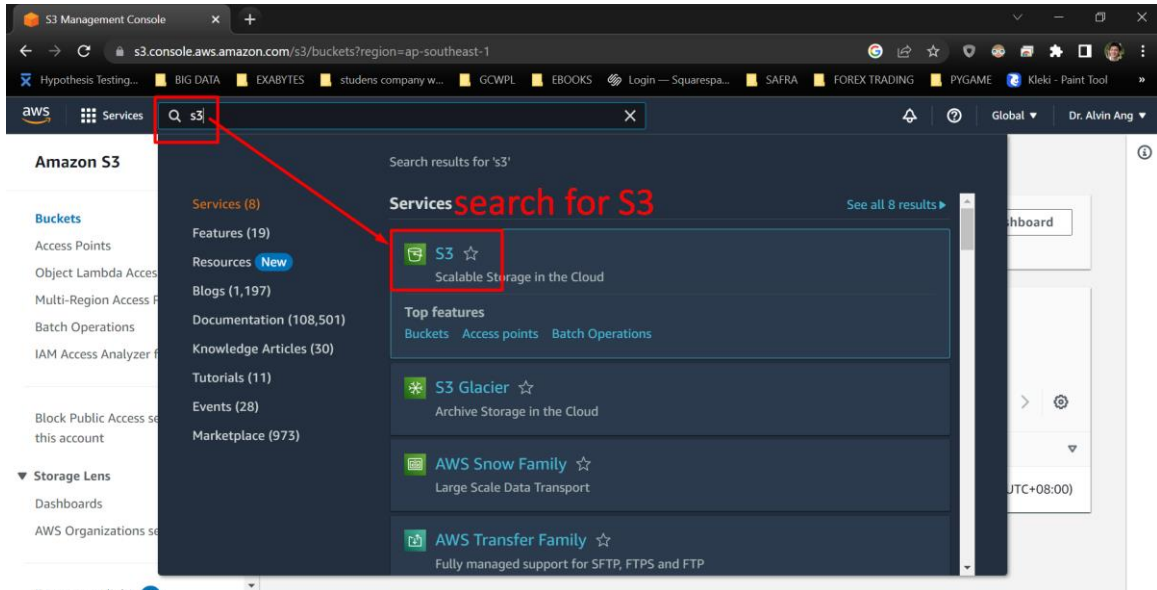
For Databricks to launch clusters in your AWS account, you must create a cross-account IAM role that gives access to Databricks. [Learn more](#)

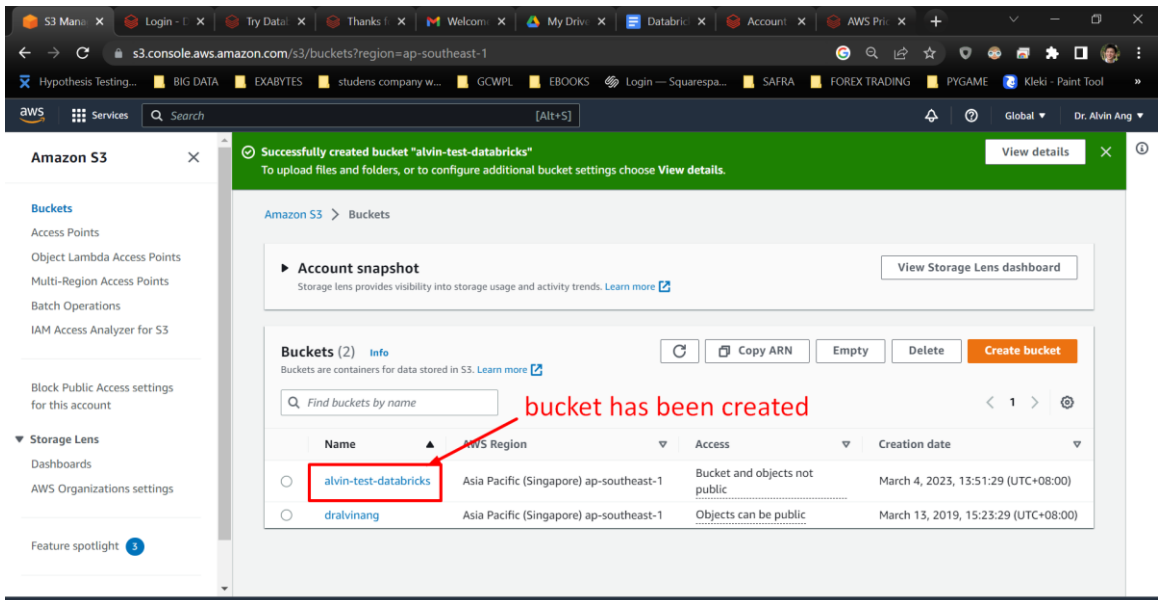
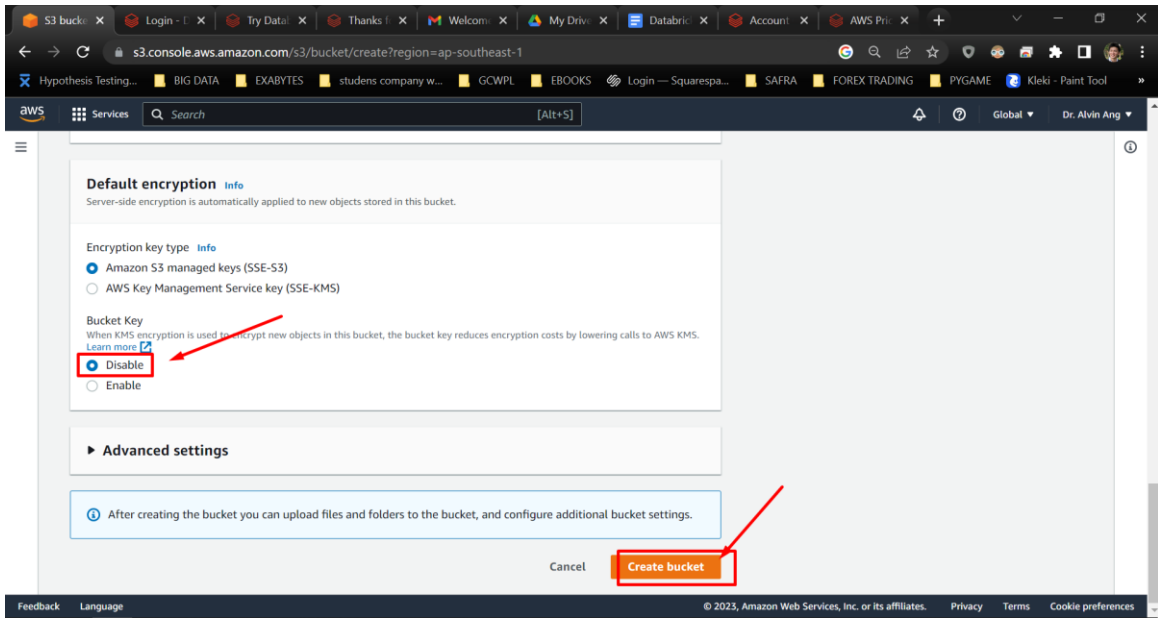
Search Search Add credential configuration

Name	Role ARN	Created
configuration-credentials-db-aws	arn:aws:iam::055975425633:role/databricks-aws-demo	today at 7:28 PM

successfully created!

IV. CREATE S3 BUCKET IN AWS

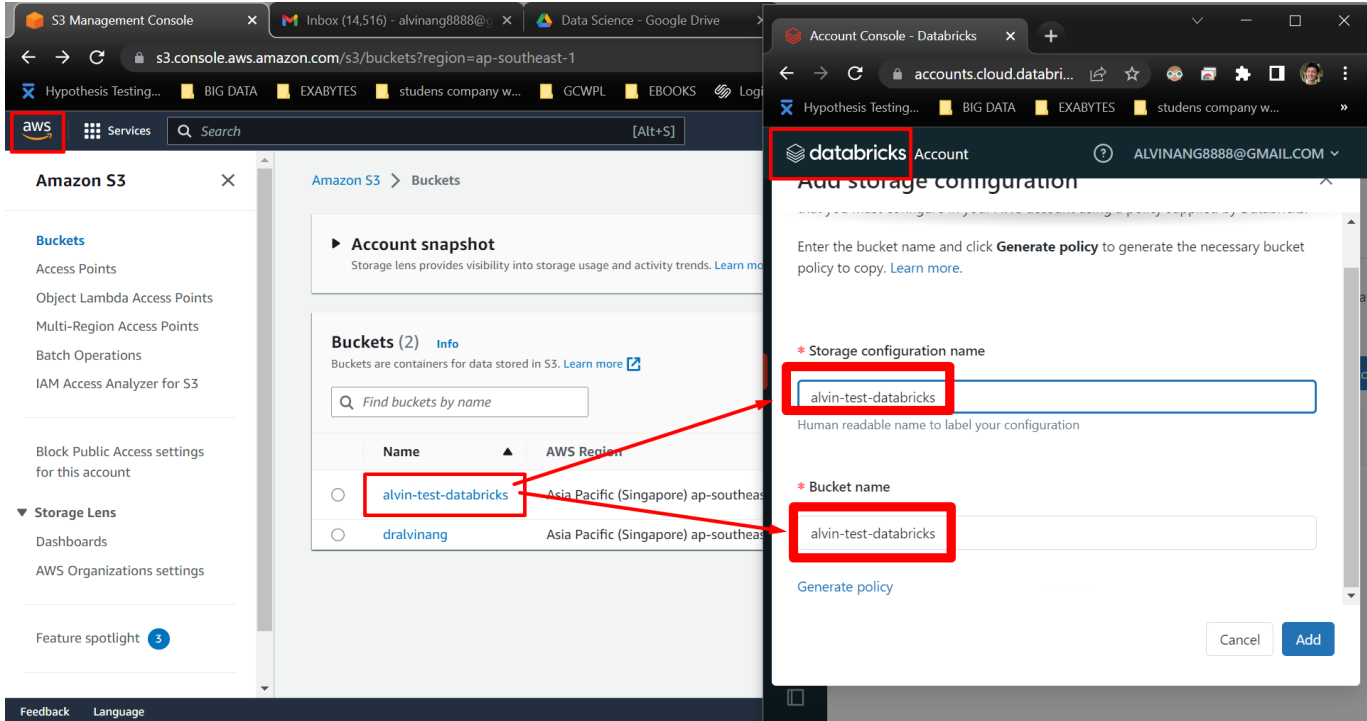




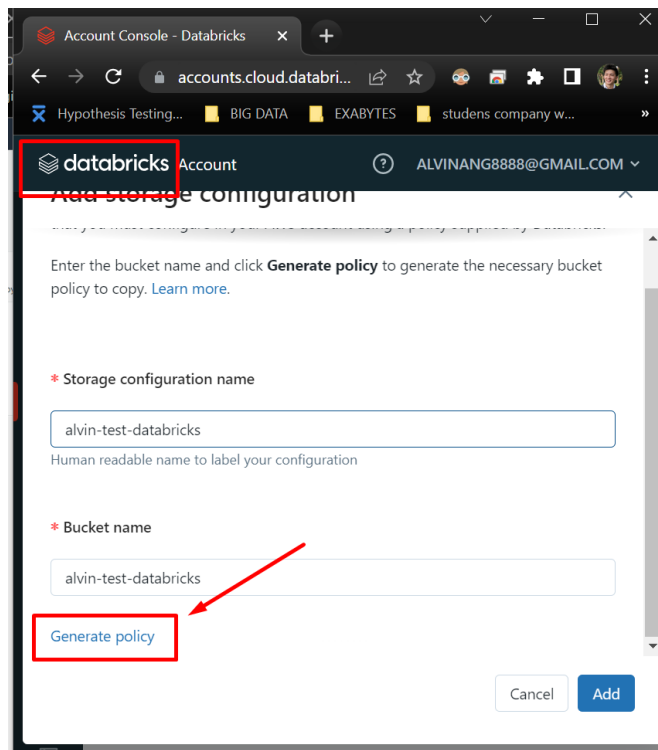
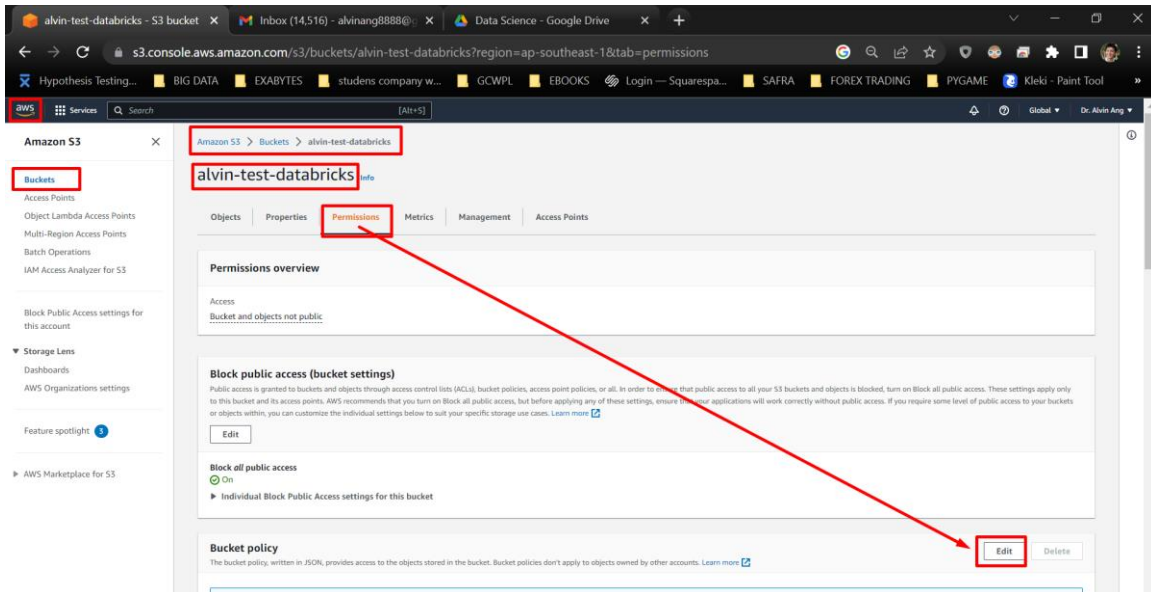
V. DATABRICKS STORAGE CONFIGURATION

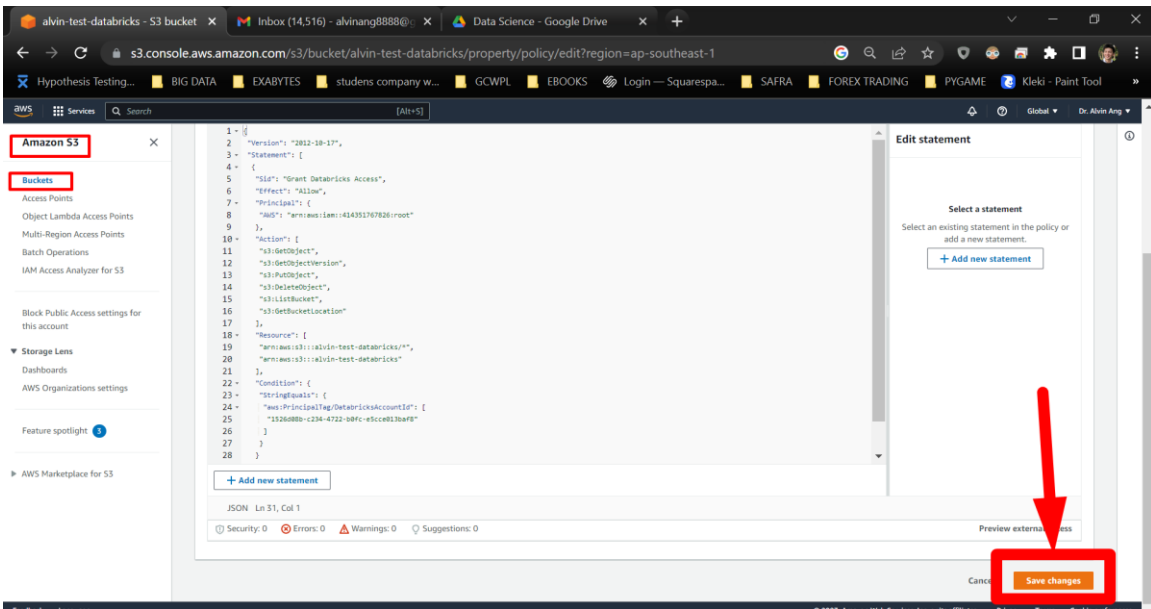
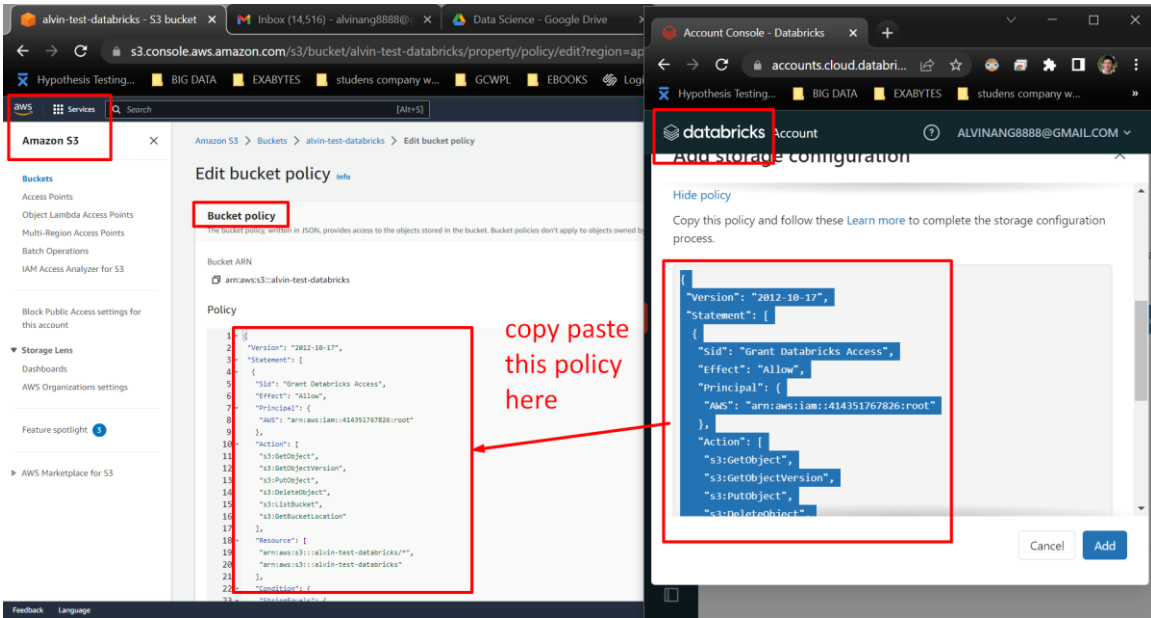
A. STORAGE CONFIGURATION NAME AND BUCKET NAME

Recall previously the BUCKET name is called 'alvin-test-databricks'



B. BUCKET POLICY





aws

Amazon S3

Successfully edited bucket policy.

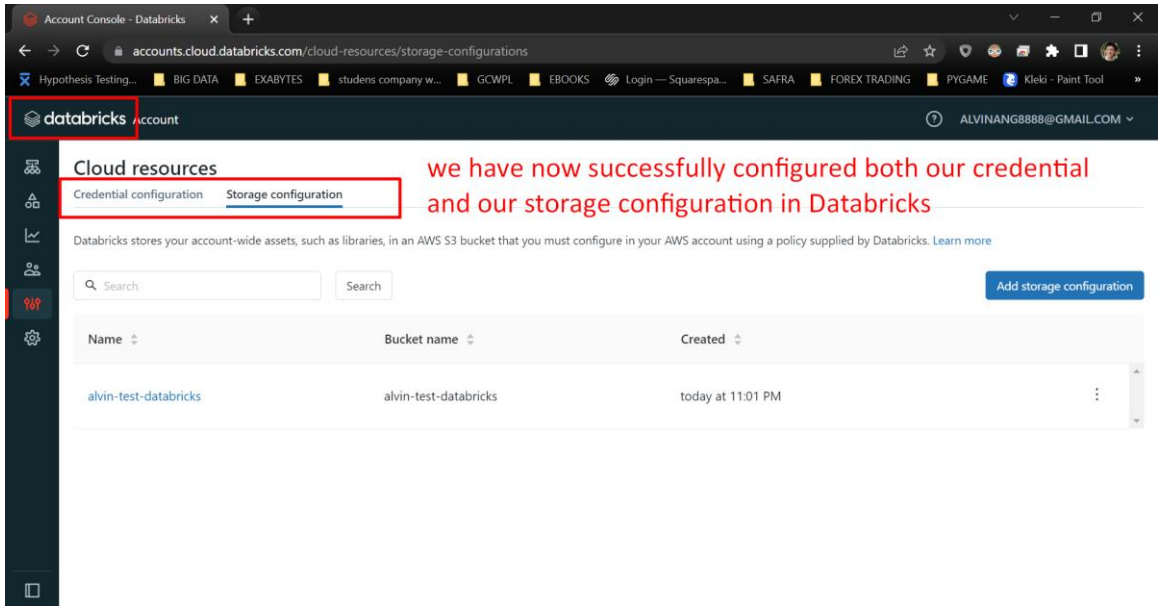
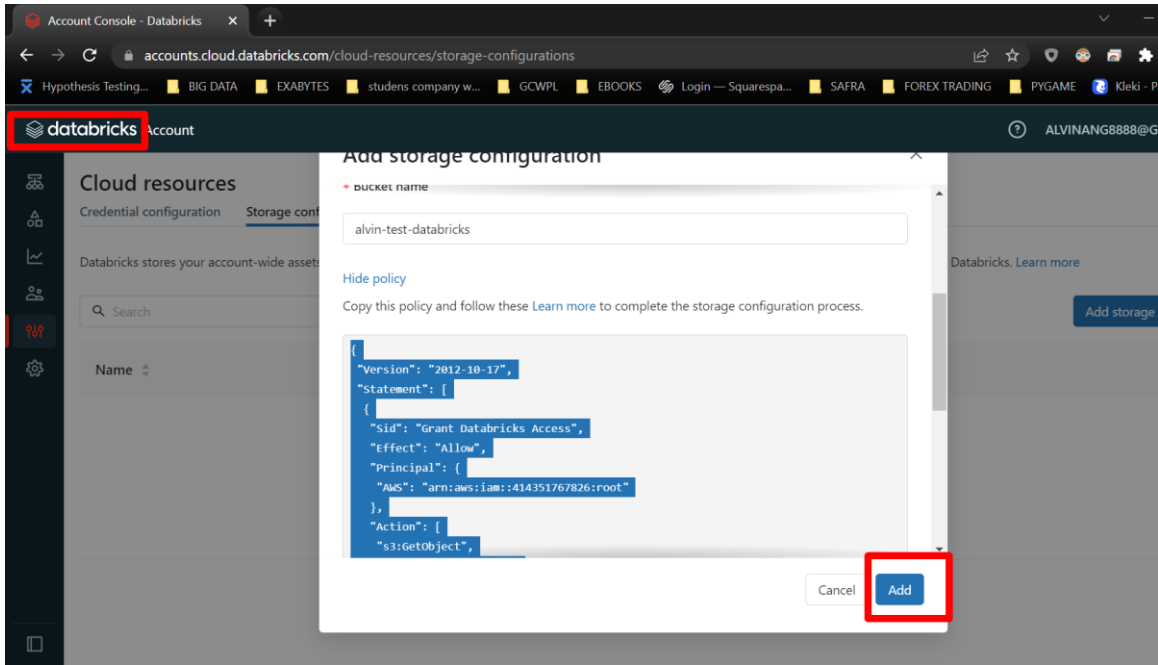
The bucket policy, written in JSON, provides access to the objects stored in the bucket. Bucket policies don't apply to objects owned by other accounts.

we have done everything we need to setup AWS already
we can actually exit but we just leave it on first
we can now move over to Databricks completely

Public access is blocked because Block Public Access settings are turned on for this bucket
To determine which settings are turned on, check your Block Public Access settings for this bucket. Learn more about using Amazon S3 Block Public Access

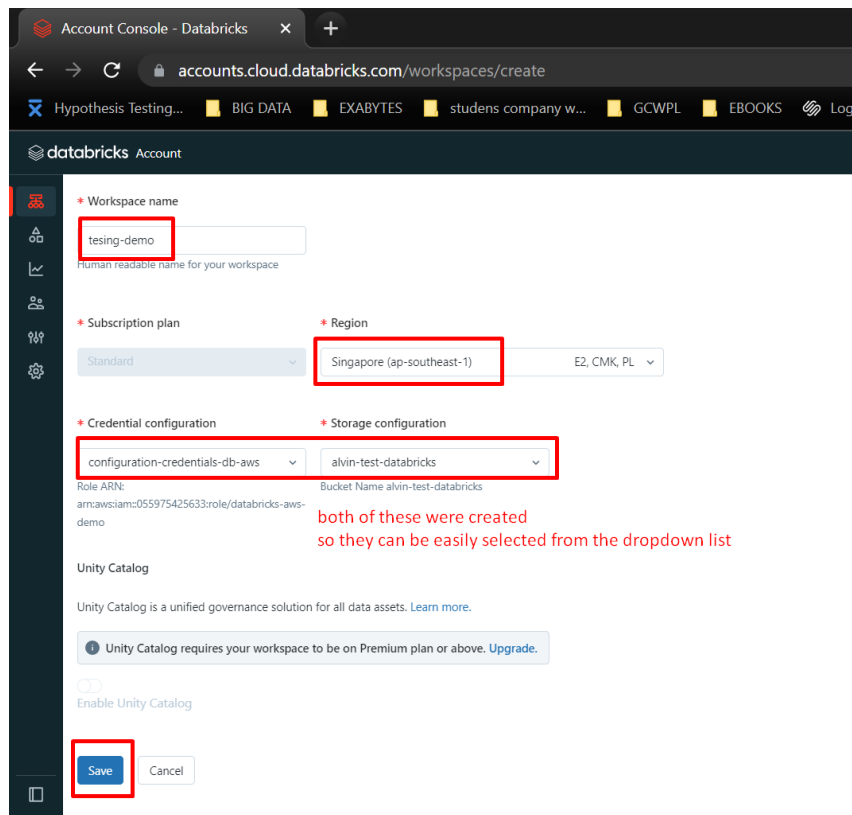
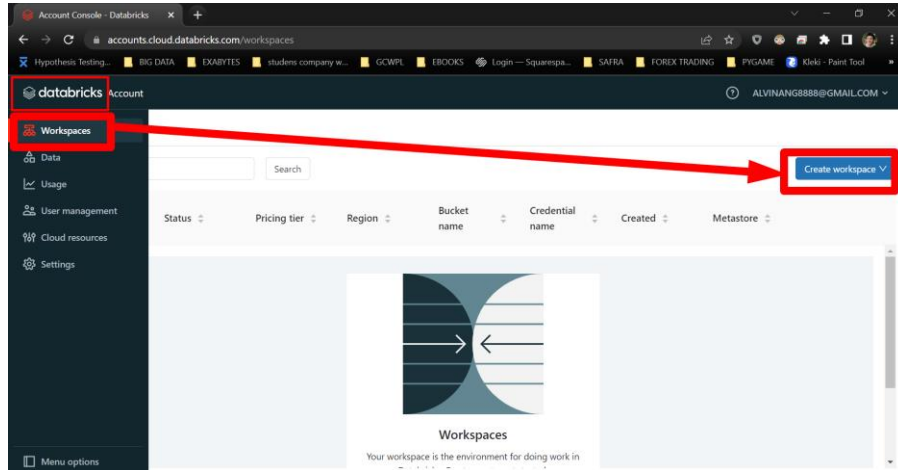
```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Sid": "Grant Databricks Access",
      "Effect": "Allow",
      "Principal": {
        "AWS": "arn:aws:iam::414351767826:root"
      },
      "Action": [
        "s3:GetObject",
        "s3:GetObjectVersion"
      ]
    }
  ]
}
```

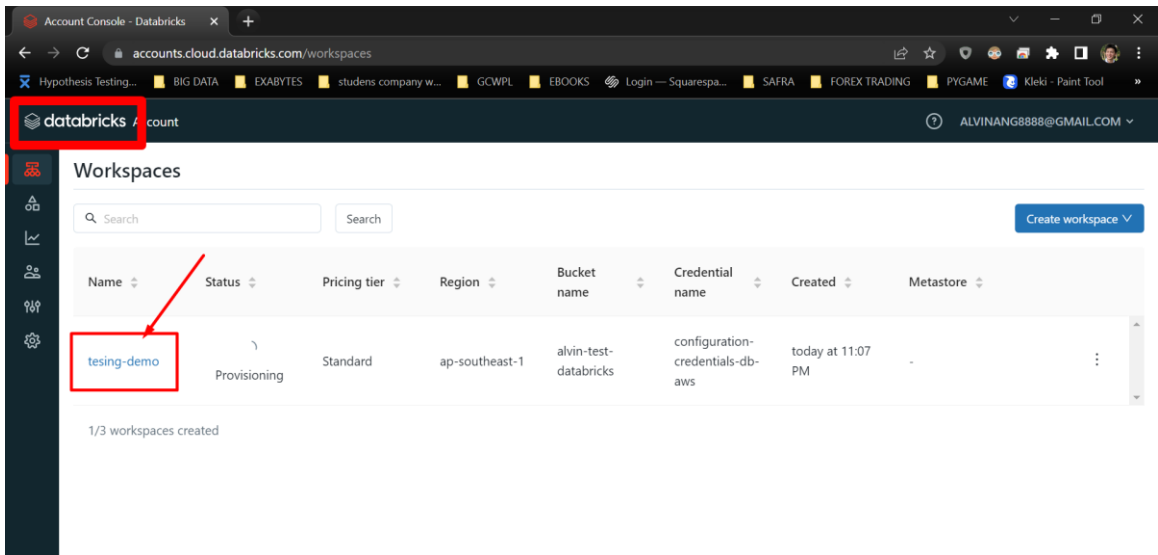
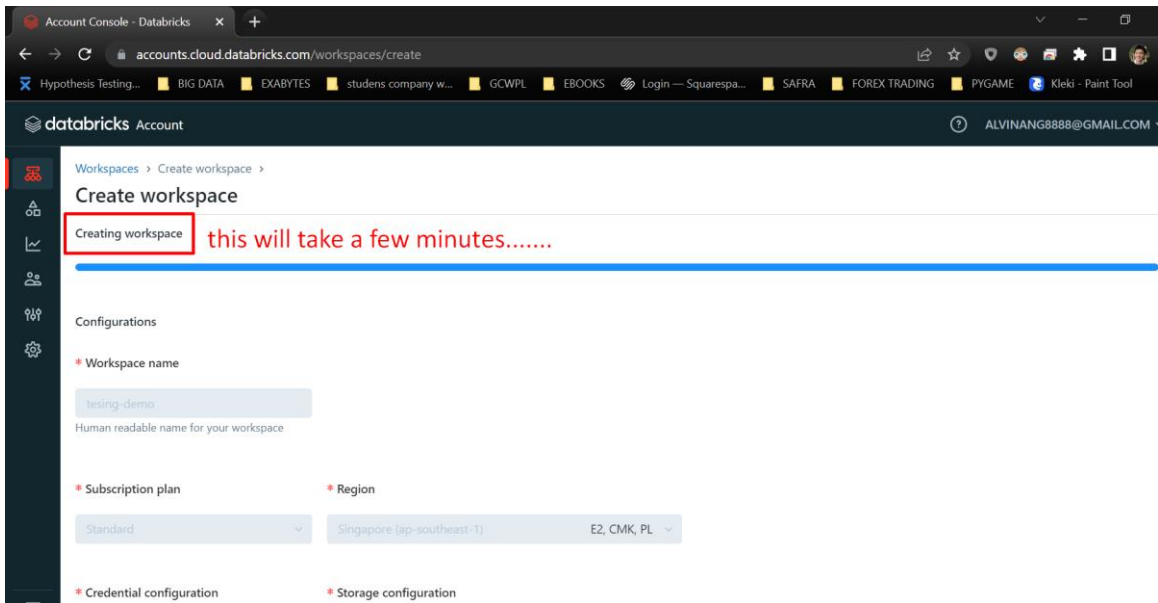
C. FINALIZE CONFIGURATIONS IN DATABRICKS

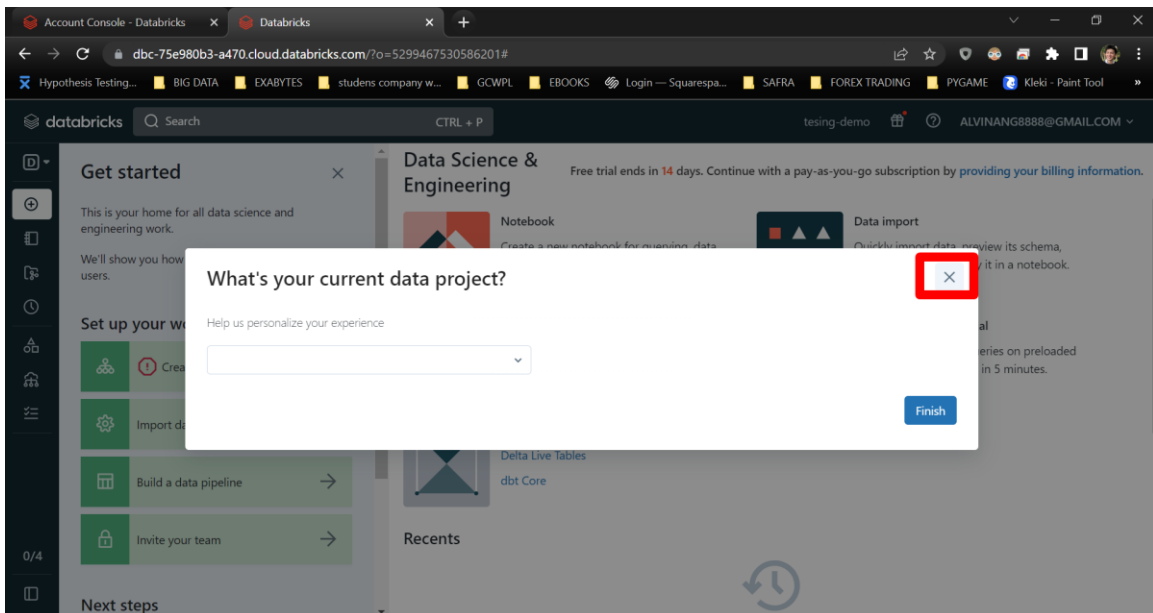
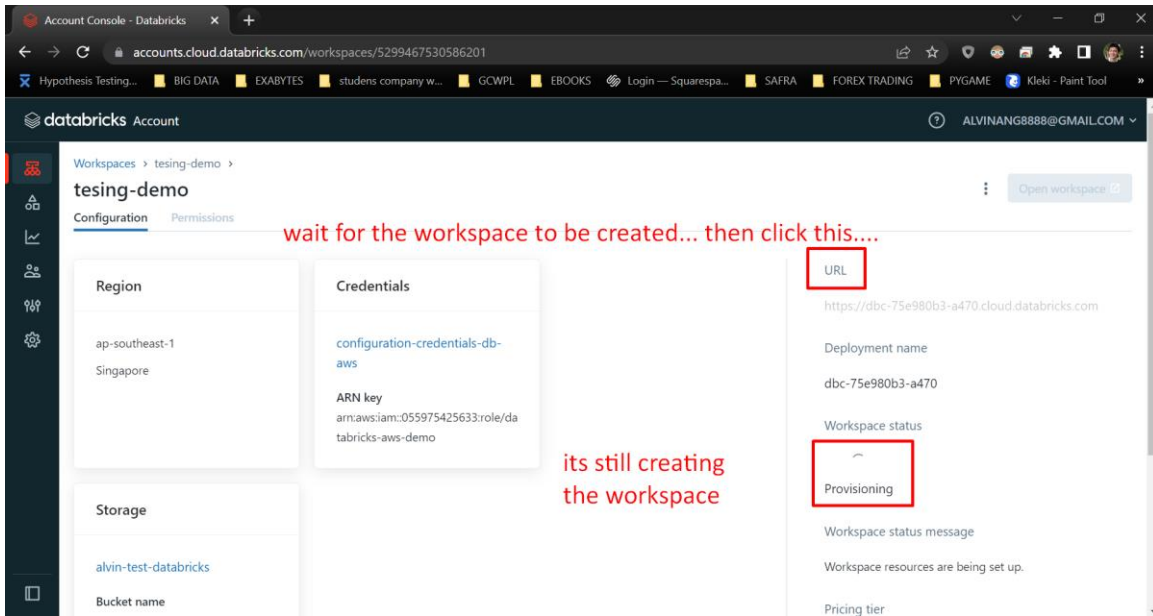


VI. CREATE WORKSPACE IN DATABRICKS

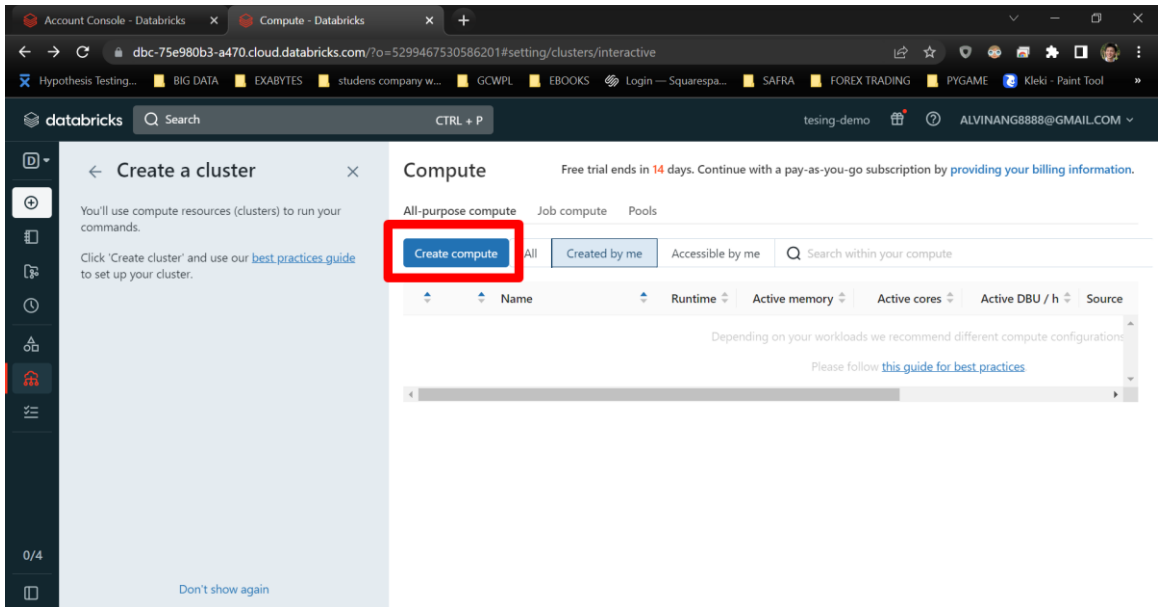
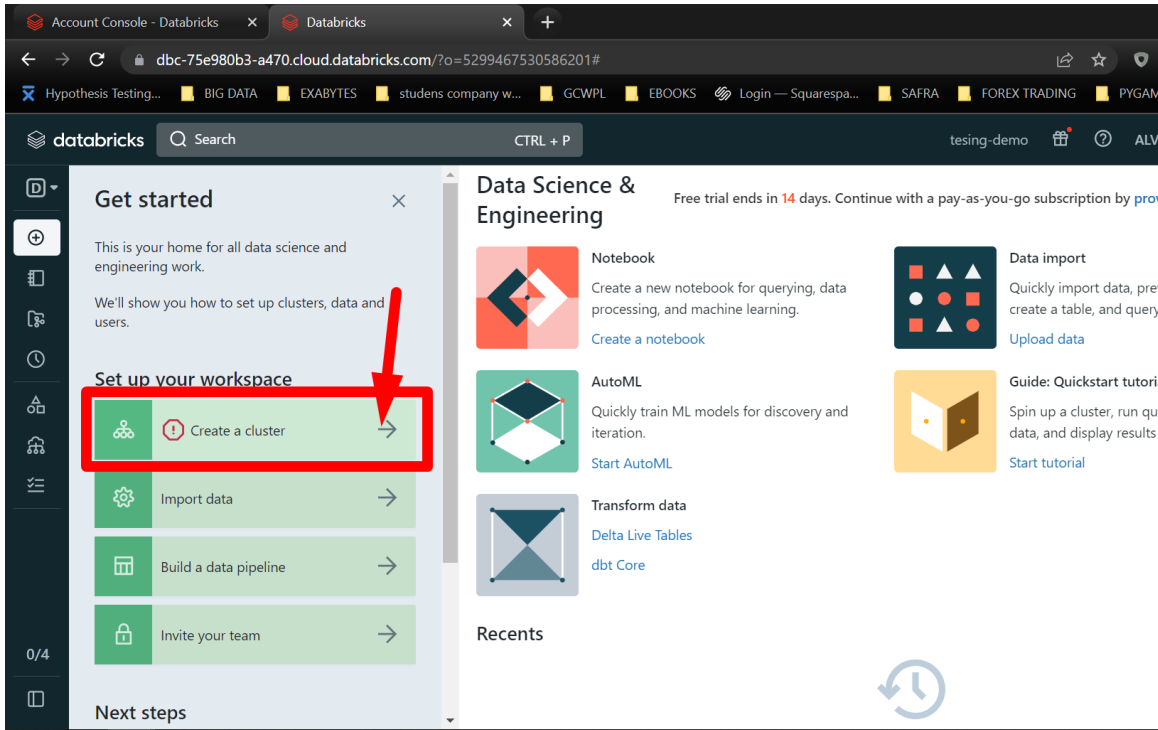
A. SETUP WORKSPACE



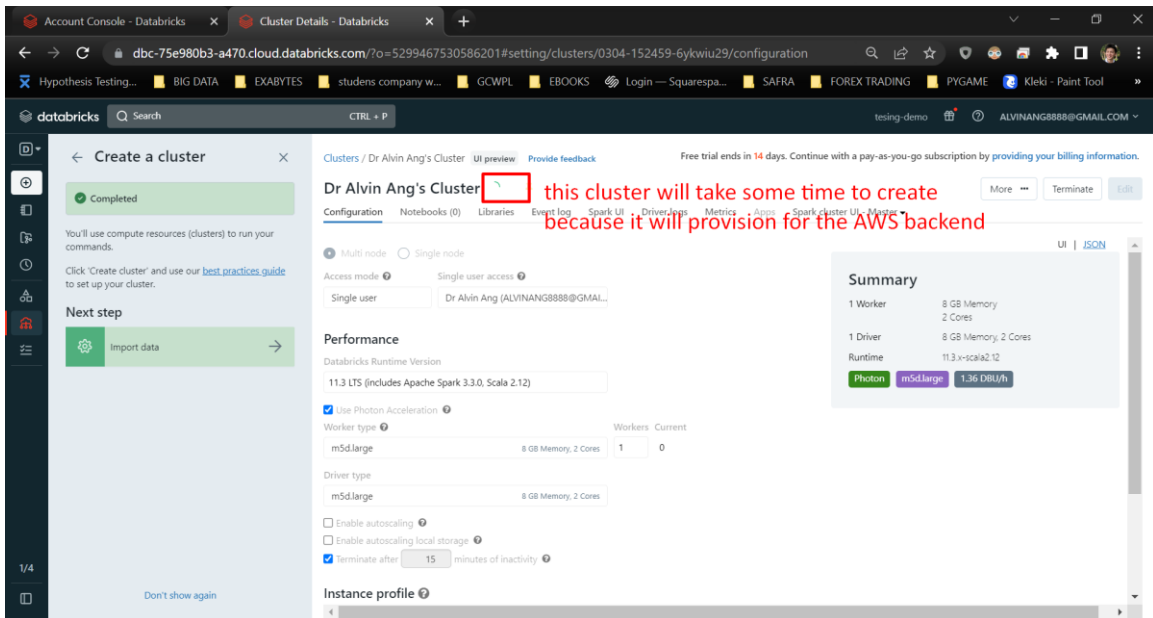
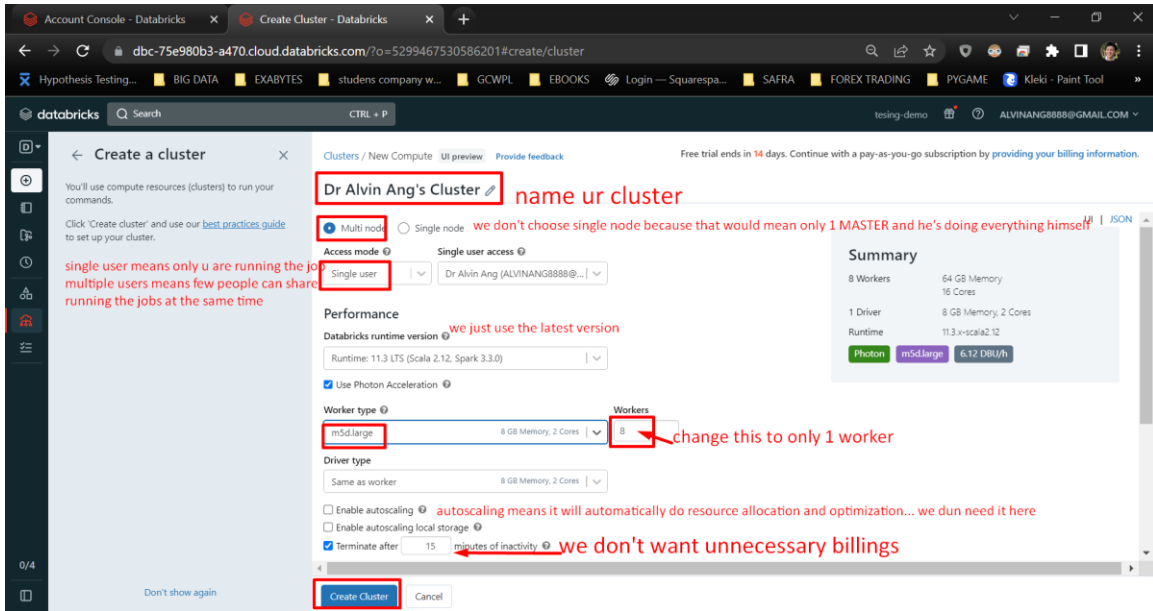




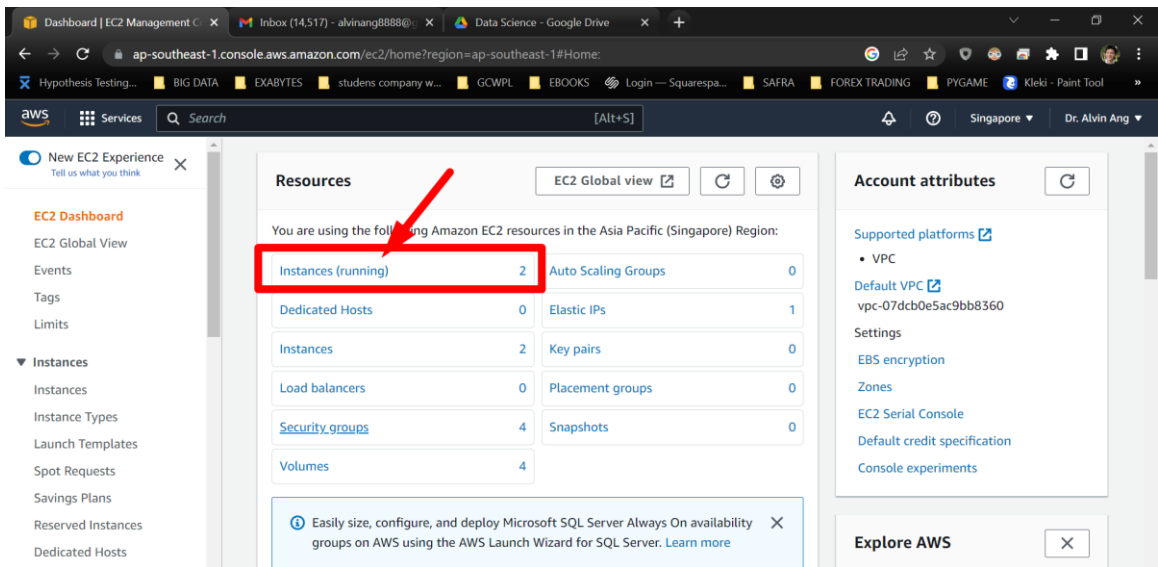
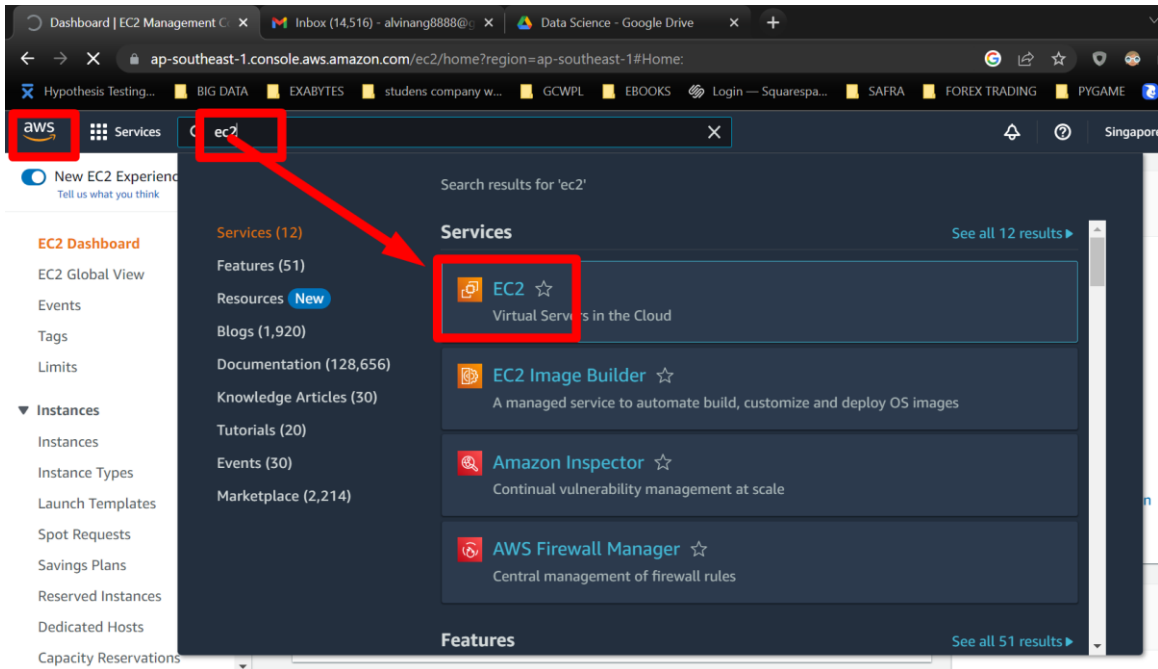
B. CREATE A CLUSTER

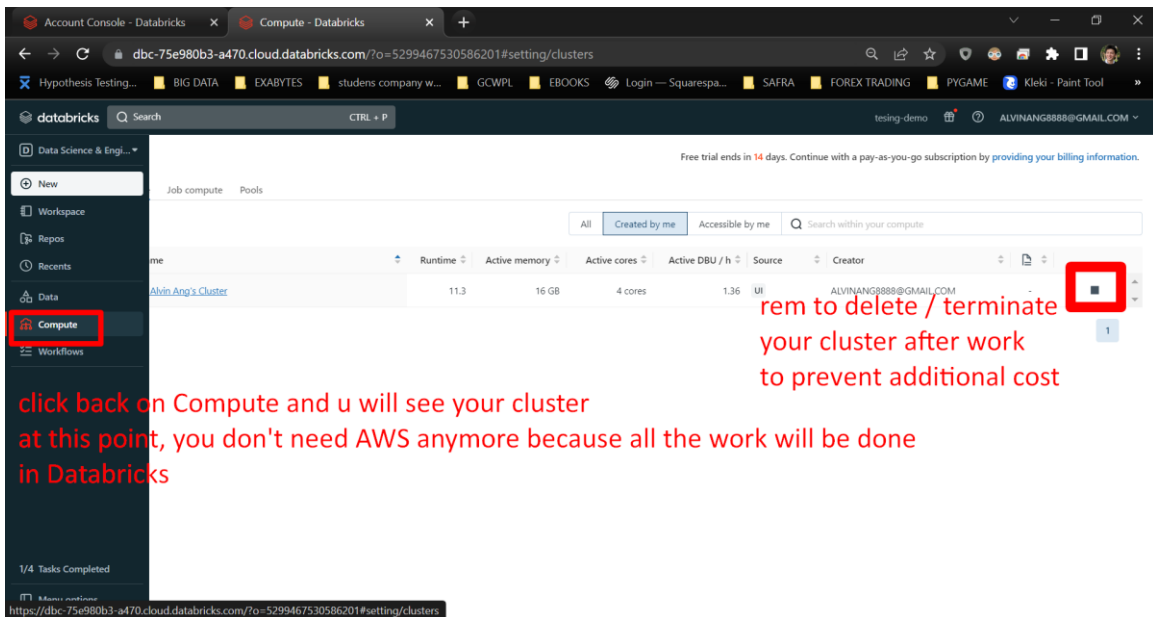
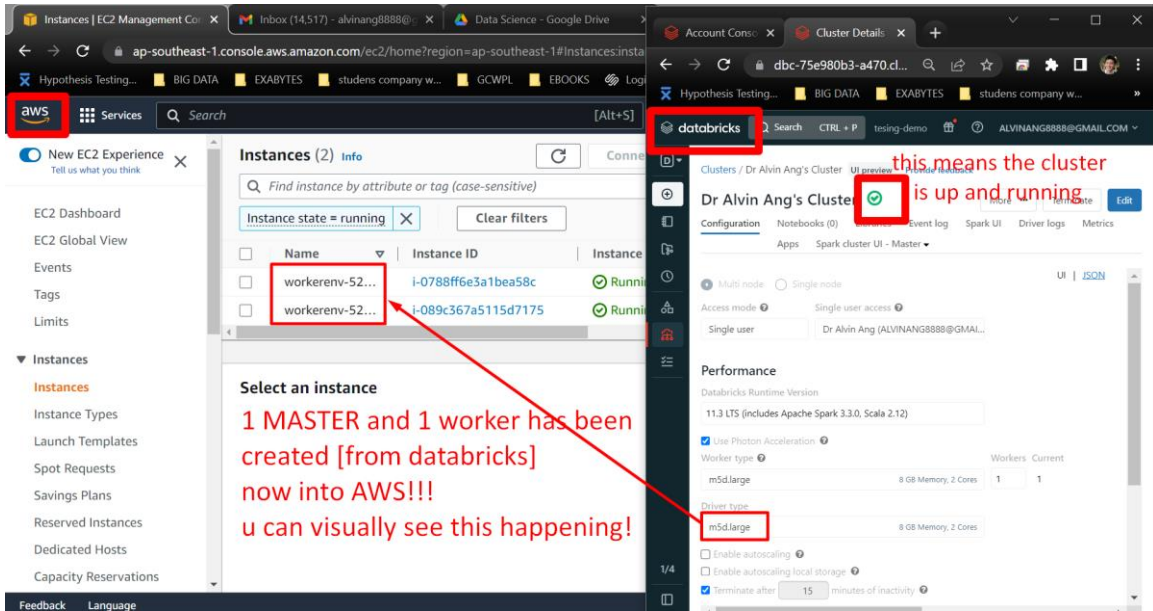


C. CONFIGURE CLUSTER



D. MEANWHILE, HEAD OVER TO AWS EC2





VII. IMPORTANT NOTE: DIFFERENCE BETWEEN ACCOUNT MODE VS WORKSPACE MODE

A. ACCOUNT MODE VS WORKSPACE MODE

The image shows two side-by-side browser windows. The left window is titled 'Account Console - Databricks' and shows the 'Account console' interface. The right window is titled 'Compute - Databricks' and shows the 'Compute workspace' interface. Both windows have a dark sidebar with various icons. Red boxes highlight the top of the sidebar in both, and red arrows point from the text 'even though their side panels look identical... THEY ARE NOT!!' to the sidebar icons in both windows.

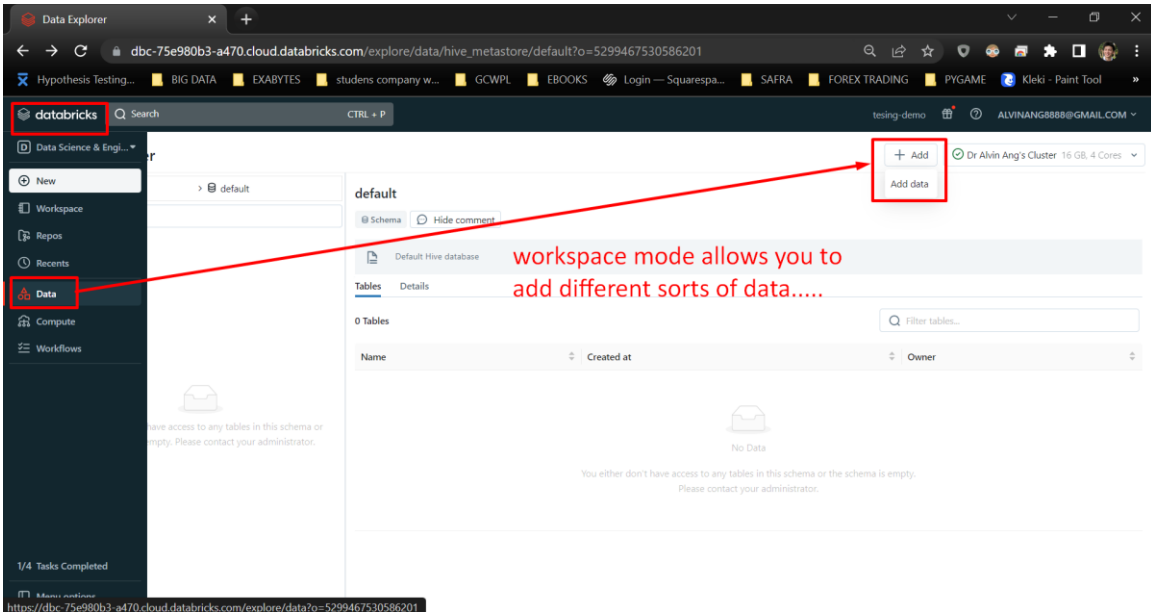
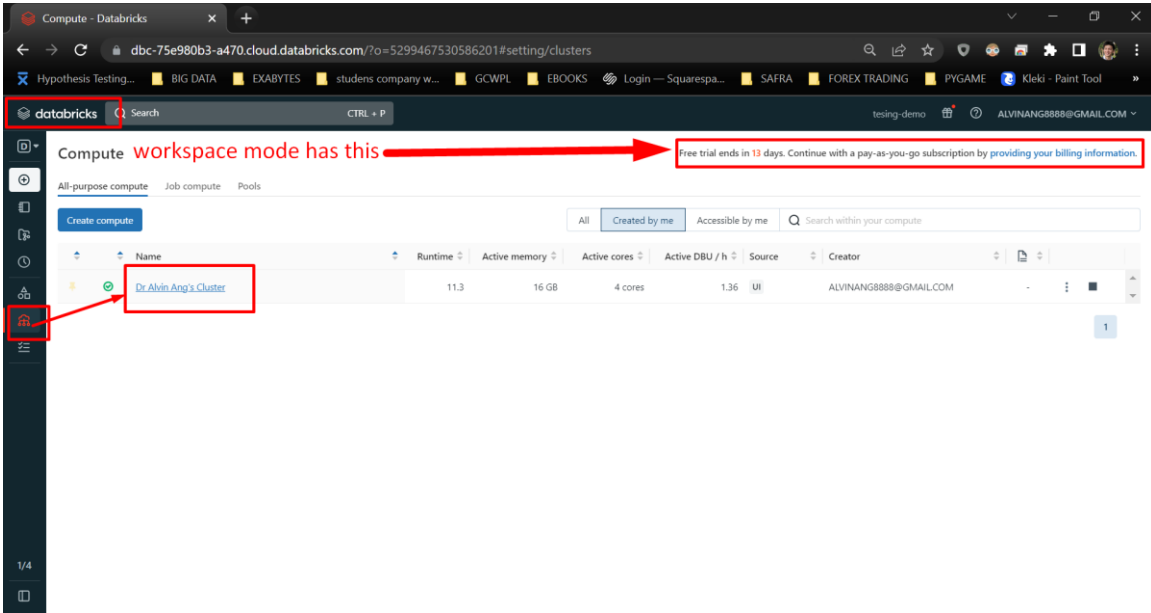
Account console account mode

Compute workspace mode

even though their side panels look identical... THEY ARE NOT!!

Name	Runtime	Active memory	Active cores	Active DBU / h
Dr. Alvin Ang's Cluster	11.3	16 GB	4 cores	

1. WORKSPACE MODE



2. ACCOUNT MODE

Account Console - Databricks

accounts.cloud.databricks.com/cloud-resources/credential-configurations

databricks Account ALVINANG8888@GMAIL.COM

Cloud resources [like if you have multiple workspaces or multiple clusters]

Credential configuration Storage configuration

For Databricks to launch clusters on your AWS account, you must create a cross-account role that gives access to Databricks. Learn more

Search Search Add credential configuration

Name	Role ARN	Created
configuration-credentials-db-aws	arn:aws:iam::055975425633:role/databricks-aws-demo	yesterday at 7:28 PM

Account Console - Databricks

accounts.cloud.databricks.com/data

databricks Account ALVINANG8888@GMAIL.COM

Data

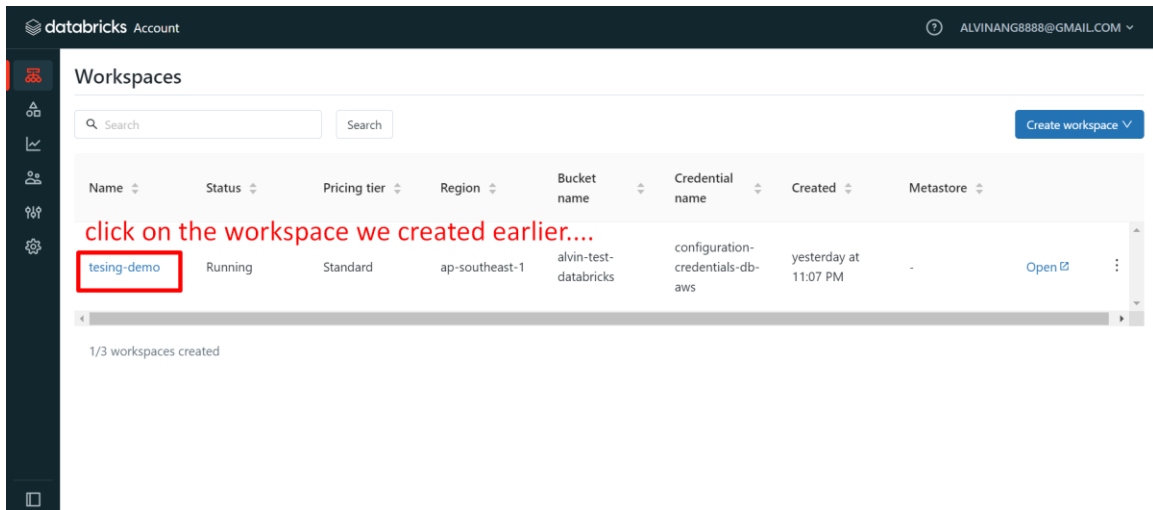
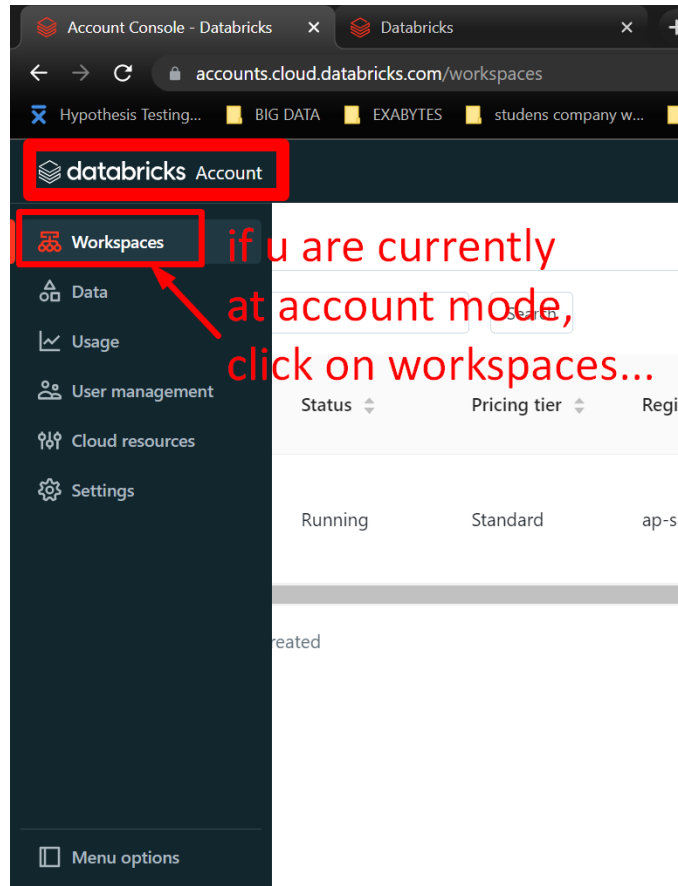
Metastores only organize them in terms of "metadata".....

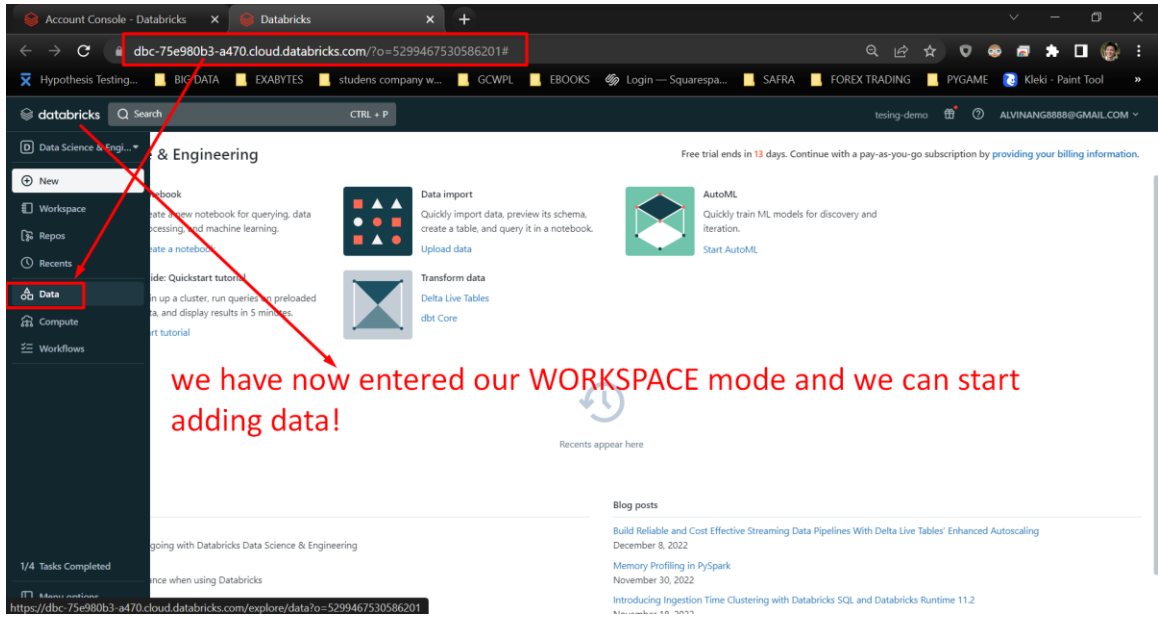
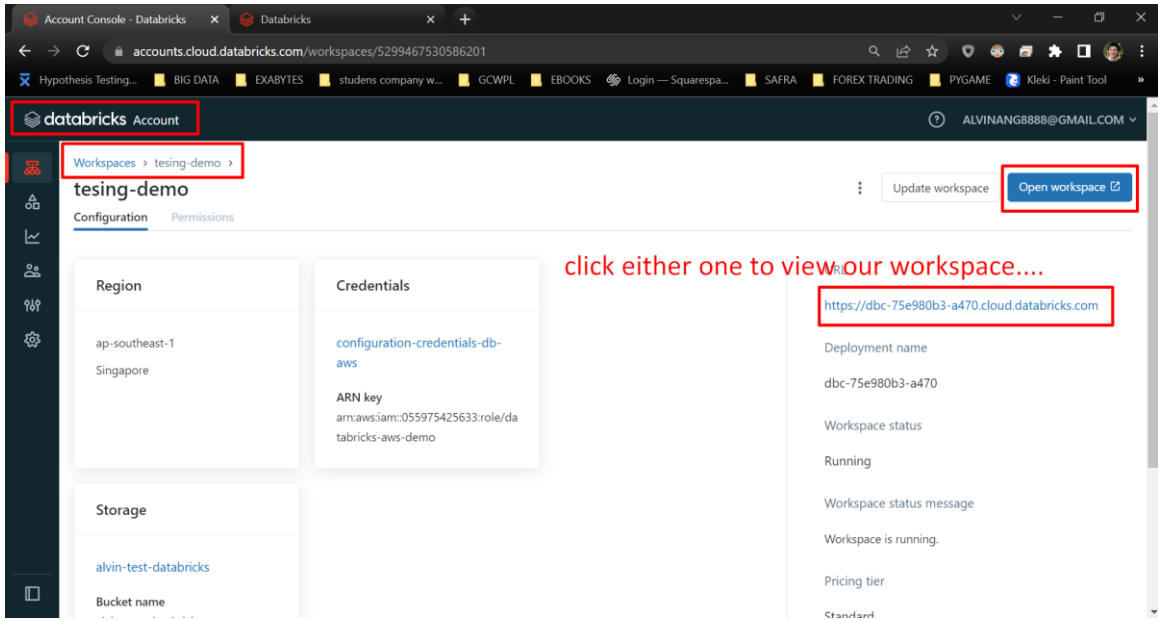
A metastore is the top-level container for data in Unity Catalog. Within a metastore, Unity Catalog provides a 3-level namespace for organizing data: catalogs, schemas (also called databases), and tables / views. Learn More

Search Search Create metastore

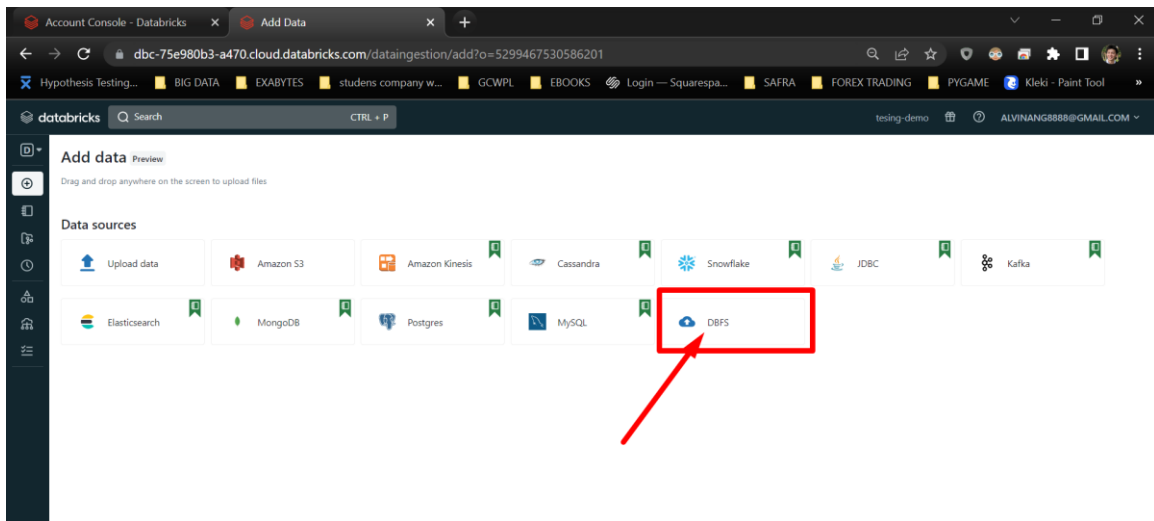
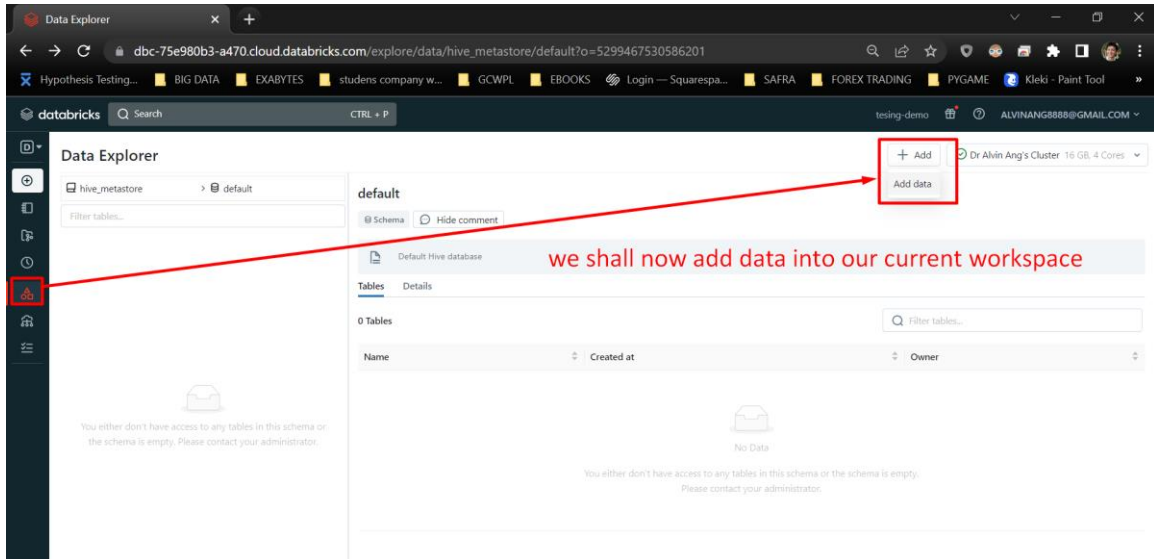
Name	Region	Path	Created at	Updated at
No metastores				

B. YOU MUST MOVE FROM ACCOUNT MODE TO WORKSPACE MODE





VIII. CREATE TABLES IN DATABRICKS



Account Console - Databricks x DBFS - Databricks

dbc-75e980b3-a470.cloud.databricks.com/?o=5299467530586201#tables/new/file

Free trial ends in 11 days. Continue with a pay-as-you-go subscription by providing your billing information.

DBFS

Upload File

DBFS Target Directory: /FileStore/tables/ (optional)

Files uploaded to /FileStore/tables/telecom_churn.csv

Files

telecom_churn.csv ✓
0.1 MB
Remove file

upload our data.....

✓File uploaded to /FileStore/tables/telecom_churn.csv

Create Table with UI

Select a Cluster to Preview the Table

Choose a cluster with which you will read and preview the data.

Cluster

Dr Alvin Ang's Cluster

select the cluster we created earlier.....

16 GB · 4 Cores · DBR 11.3 LTS · Photon · Spark 3.3.0 · Scala 2.12

Account Console - Databricks x DBFS - Databricks

dbc-75e980b3-a470.cloud.databricks.com/?o=5299467530586201#tables/new/file

Free trial ends in 13 days. Continue with a pay-as-you-go subscription by providing your billing information.

DBFS

Dr Alvin Ang's Cluster

Preview Table

Specify Table Attributes

Specify the Table Name, Database and Schema to add this to the data UI for other users to access

Table Name: telecom_churn_csv

Create in Database: default

File Type: CSV

Column Delimiter: ,

First row is header

Infer schema

Multi-line

Create table

Table Preview

Churn	AccountWeeks	ContractRenewal	DataPlan	DataUsage	CustServCalls	DayMins
0	128	1	1	2.7	1	265.1
0	107	1	1	3.7	1	161.6
0	137	1	0	0	0	243.4
0	84	0	0	0	2	299.4
0	75	0	0	0	3	166.7
0	118	0	0	0	0	223.4

after the Table has been created....

we can now preview the Table.....

Churn	AccountWeeks	ContractRenewal	DataPlan	DataUsage	CustServCalls	DayMins	DayCalls	MonthlyCharge	OverageFee	...
0	128	1	1	2.7	1	265.1	110	89	9.87	10
0	107	1	1	3.7	1	161.6	123	82	9.78	13
0	137	1	0	0	0	243.4	114	52	6.06	12
0	84	0	0	0	2	299.4	71	57	3.1	61
0	75	0	0	0	3	166.7	113	41	7.42	10
0	118	0	0	0	0	223.4	98	57	11.03	6
0	121	1	1	2.03	3	218.2	88	87.3	17.43	7
0	147	0	0	0	0	157	79	36	5.16	7

IX. CREATE NOTEBOOK

The screenshot shows the Databricks Data Explorer interface. The left sidebar has a 'New' button highlighted with a red box. A dropdown menu is open, showing 'Notebook' highlighted with a red box. A red arrow points to the 'Notebook' option with the text 'create a new notebook'. A red text overlay says 'note we are currently in WORKSPACE mode...'. The main area displays a table with columns: Churn, AccountWeeks, ContractRenewal, DataPlan, DataUsage, CustServCalls, DayMins, DayCalls, MonthlyCharge, and OverageFee. The table contains 10 rows of data.

Churn	AccountWeeks	ContractRenewal	DataPlan	DataUsage	CustServCalls	DayMins	DayCalls	MonthlyCharge	OverageFee
0	128	1	1	2.7	1	265.1	110	89	9.87
0	107	1	1	3.7	1	161.6	123	82	9.78
0	137	1	0	0	0	243.4	114	52	6.06
0	84	0	0	0	2	299.4	71	57	3.1
0	75	0	0	0	3	166.7	113	41	7.42
0	118	0	0	0	0	223.4	98	57	11.03
0	121	1	1	2.03	3	218.2	88	87.3	17.43
0	147	0	0	0	0	157	79	36	5.16

The screenshot shows the 'Create Notebook' dialog box. The 'Name' field contains 'testing_notebook'. The 'Default Language' dropdown is set to 'Python'. The 'Cluster' dropdown is set to 'Dr Alvin Ang's Cluster'. The 'Create' button is highlighted with a red box.

Create Notebook

Name: testing_notebook

Default Language: Python

Cluster: Dr Alvin Ang's Cluster

Buttons: Cancel, Create

A. ATTACH / DETACH CLUSTER...

The screenshot shows the Databricks web interface. At the top, there are browser tabs for 'Account Console - Databricks' and 'testing_notebook - Databricks'. The address bar shows the URL 'dbc-75e980b3-a470.cloud.databricks.com/#notebook/447859306851712/command/447859306851713'. The main interface includes a search bar, a 'testing_notebook' title, and a 'Run all' button. A dropdown menu is open, showing 'Dr Alvin Ang's Cluster' with a red box around it. The menu lists 'Runtime' (D8R 11.3 LTS - Spark 3.3.0 - Scala 2.12), 'Driver' (m5.xlarge - 8 GB - 2 Cores), and 'Worker' (m5.xlarge - 8 GB - 2 Cores). A red arrow points from the text below to the 'Detach & re-attach' option in the menu. The notebook content shows a code cell with the text 'our jupyter notebook environment!' and a red annotation 'our jupyter notebook environment!'.

our jupyter notebook environment!

our jupyter notebook environment!

up to this point, we have successfully setup our CLUSTER you can detach/attach your CLUSTER here....

B. START SPARK SESSION

The screenshot shows a Databricks notebook interface with three commands executed. The first command imports pyspark and SparkSession. The second command creates a SparkSession named 'spark'. The third command displays the 'spark' object, showing its class and context.

```
Cmd 1
1 import pyspark
2 from pyspark.sql import SparkSession

Command took 0.08 seconds -- by ALVINANG8888@GMAIL.COM at 3/5/2023, 2:26:16 PM on Dr Alvin Ang's Cluster

Cmd 2
1 spark = SparkSession.builder.appName("GetStarted").getOrCreate()

Command took 0.10 seconds -- by ALVINANG8888@GMAIL.COM at 3/5/2023, 2:27:50 PM on Dr Alvin Ang's Cluster

Cmd 3
1 spark

SparkSession - hive
SparkContext
Spark UI
Version
v3.3.0
Master
spark://10.185.249.16:7077
AppName
Databricks Shell
```

C. TEST SOME CODE....

The screenshot shows a Databricks notebook interface. The notebook is titled "testing_notebook" and is running Python code. The first cell contains the following code, which is highlighted with a red box:

```
df=sqlContext.sql("SELECT * FROM telecom_churn_csv")
```

Below the code, the output shows the schema of the DataFrame:

```
df: pyspark.sql.dataframe.DataFrame
  Churn: string
  AccountWeeks: string
  ContractRenewal: string
  DataPlan: string
  DataUsage: string
  CustServCalls: string
  DayMins: string
  DayCalls: string
  MonthlyCharge: string
  OverageFee: string
  RoamMins: string
```

The command took 0.32 seconds to execute. The second cell contains the following code, also highlighted with a red box:

```
df.show(5)
```

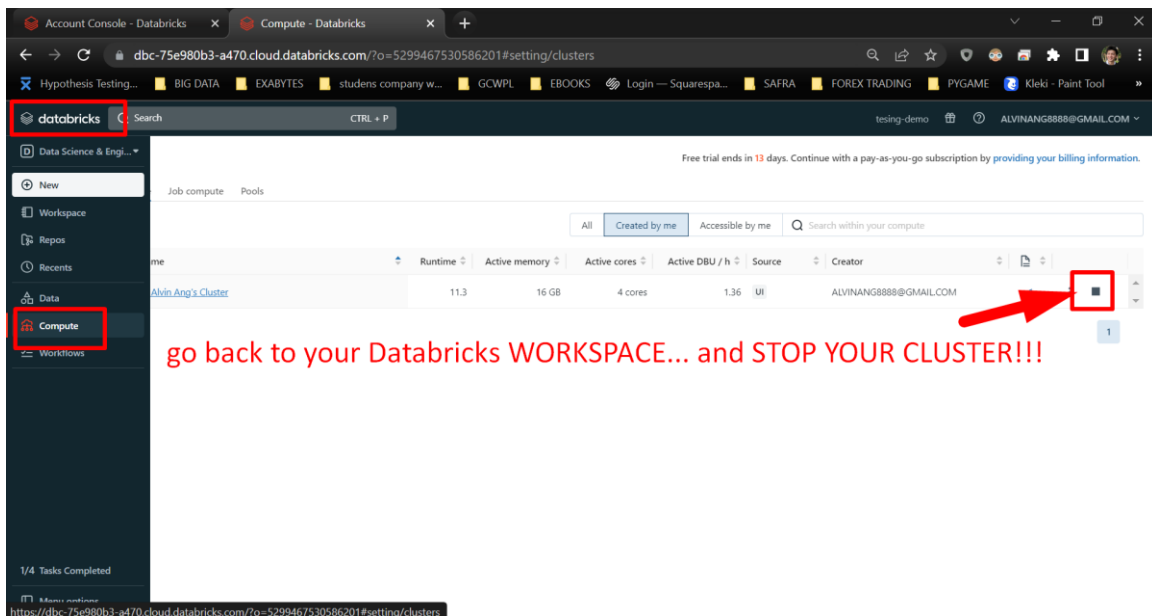
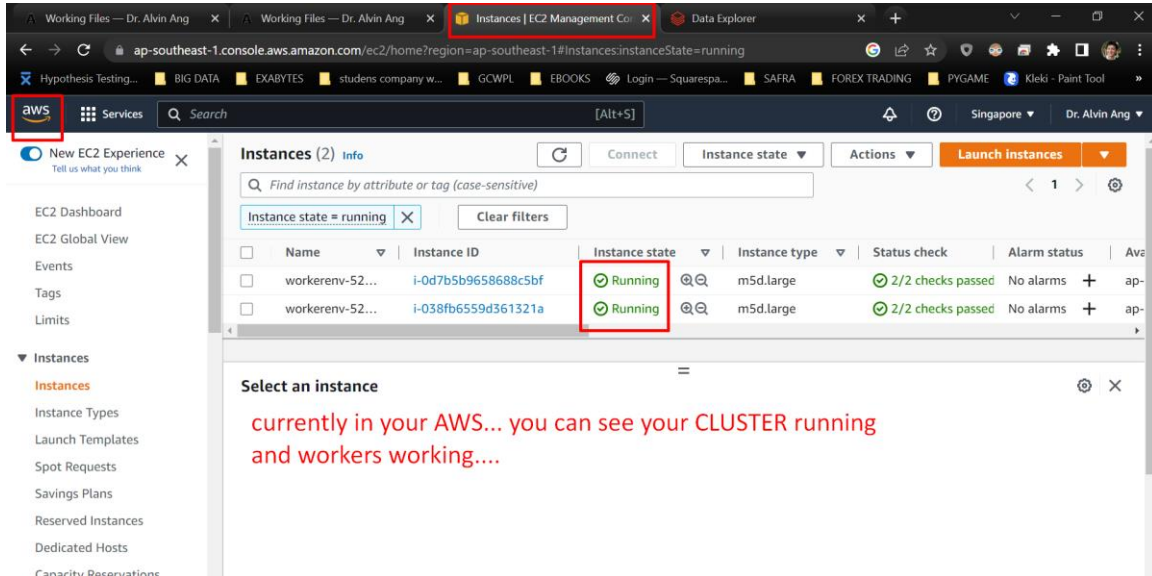
The output of the second cell shows the first five rows of the data, displayed in a table format:

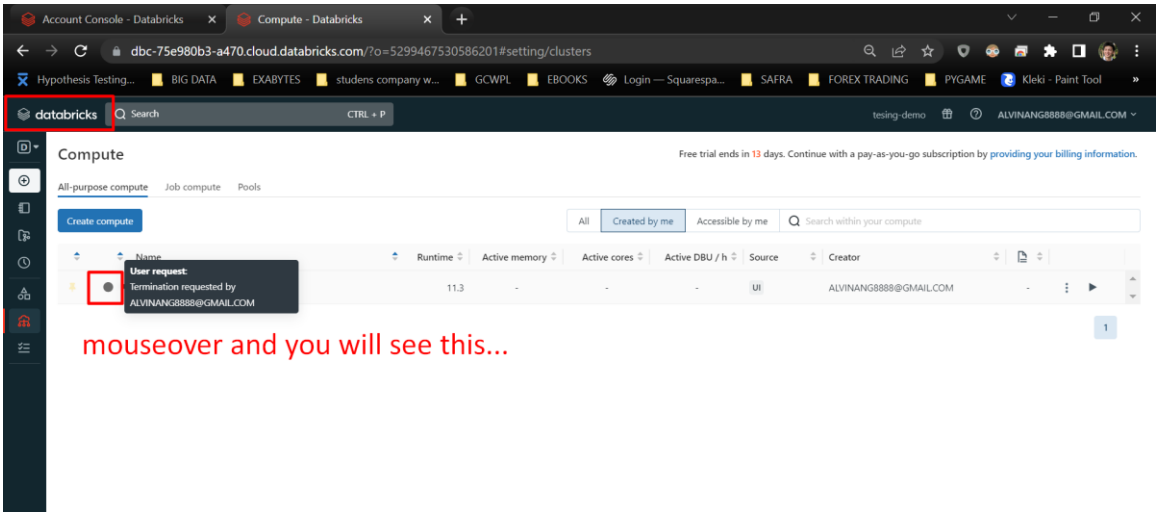
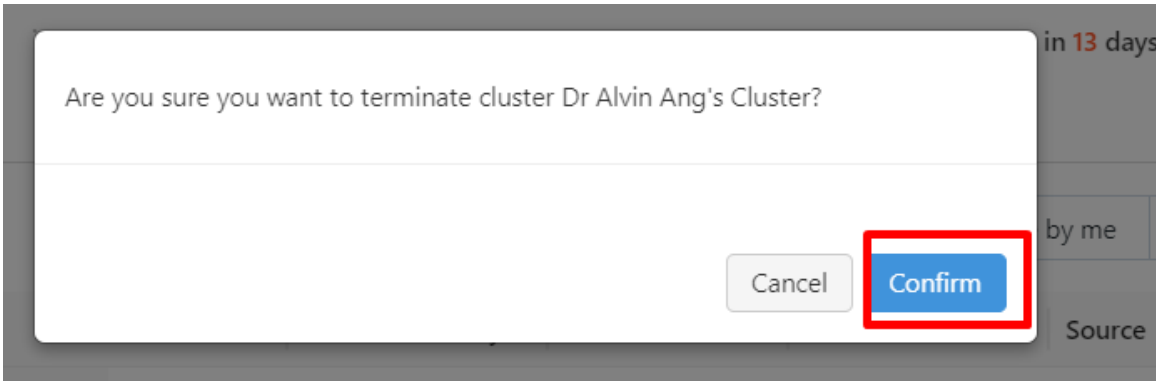
Churn	AccountWeeks	ContractRenewal	DataPlan	DataUsage	CustServCalls	DayMins	DayCalls	MonthlyCharge	OverageFee	RoamMins
0	128	1	1	2.7	1	265.1	110	89	9.87	10
0	107	1	1	3.7	1	161.6	123	82	9.78	13.7
0	137	1	0	0	0	243.4	114	52	6.06	12.2
0	84	0	0	0	2	299.4	71	57	3.1	6.6
0	75	0	0	0	3	166.7	113	41	7.42	10.1

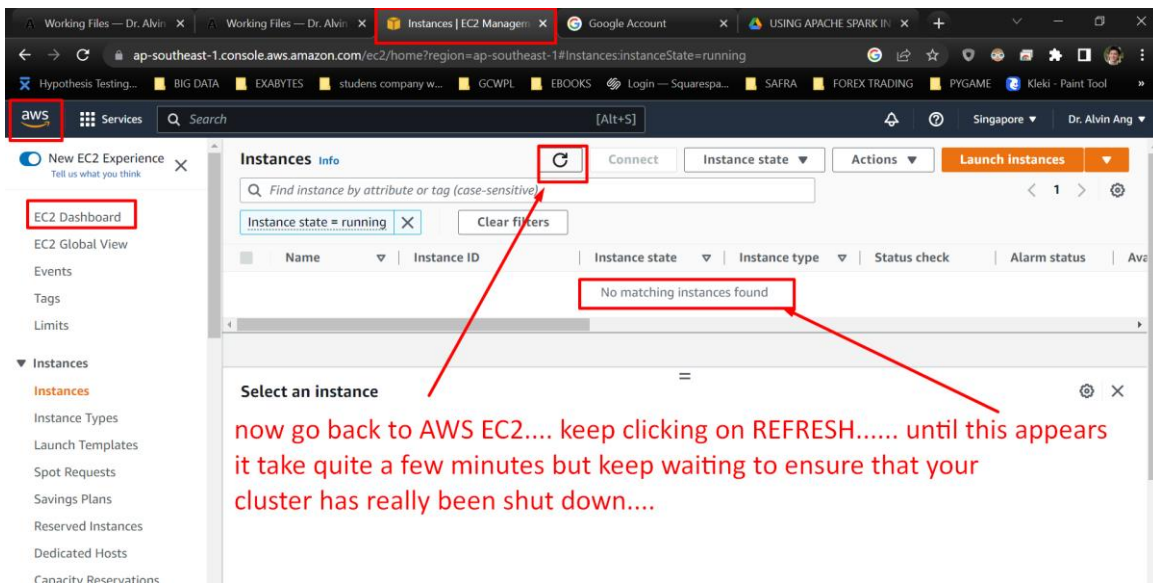
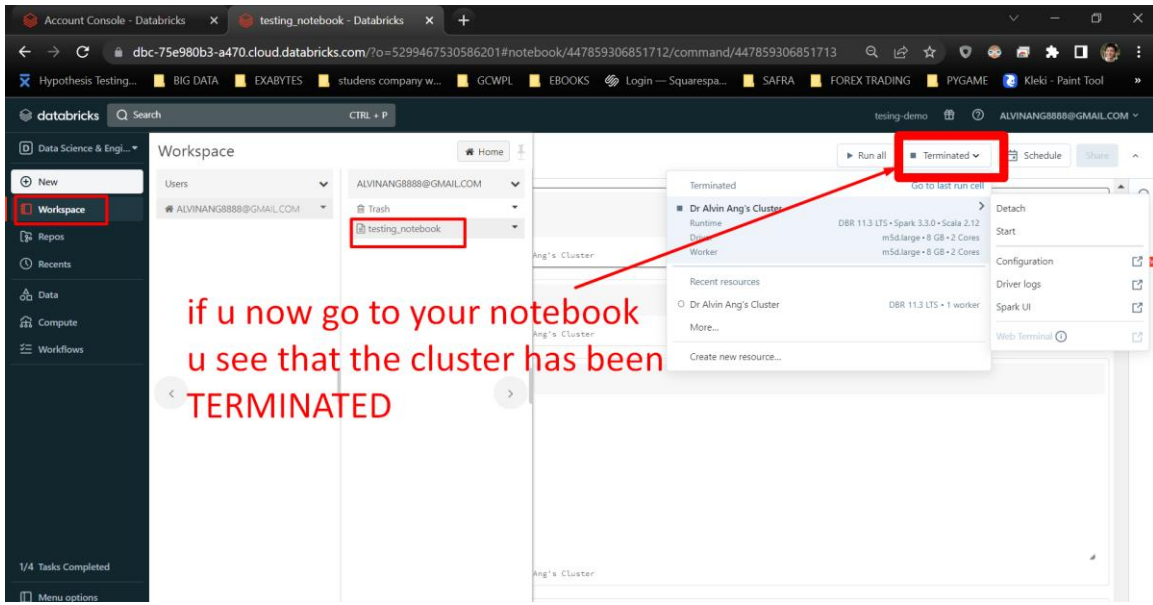
The text "IT WORKS!!!!" is written in red in the center of the notebook output area.

X. VERY IMPORTANT: SHUT DOWN YOUR CLUSTER!!!

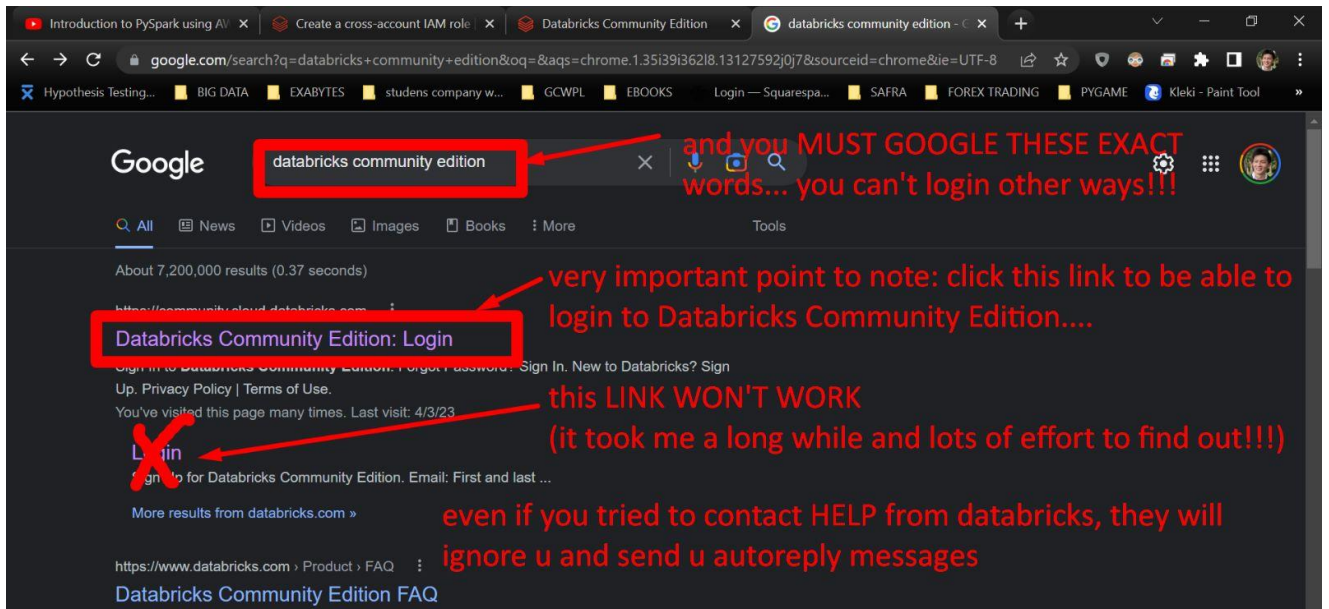
To prevent further charges...



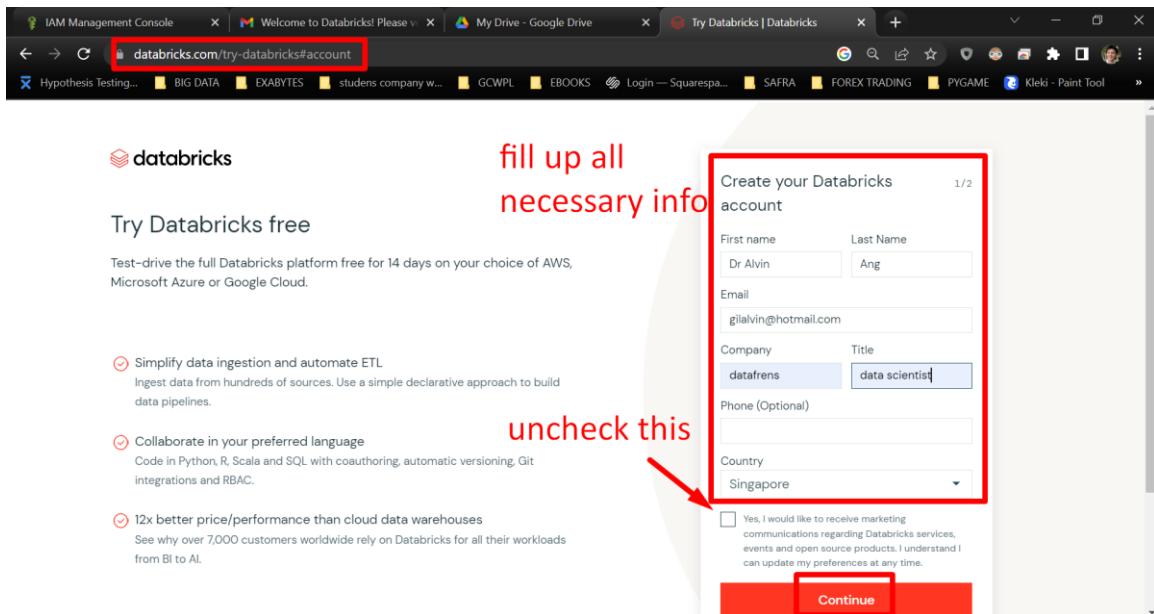




XI. APPENDIX I: SIGNING UP WITH DATABRICK COMMUNITY EDITION [FREE FOREVER BUT WE WON'T BE USING THIS OPTION IN THIS MANUSCRIPT]



<https://www.databricks.com/try-databricks#account>





do NOT choose any of these options
or they will log you out forever after 14 days!!!

Try Databricks free

Test-drive the full Databricks platform free for 14 days on your choice of AWS, Microsoft Azure or Google Cloud.

'one of my accounts was logged out forever...wasted my gmail account...'

- ✔ Simplify data ingestion and automate ETL
Ingest data from hundreds of sources. Use a simple declarative approach to build data pipelines.
- ✔ Collaborate in your preferred language
Code in Python, R, Scala and SQL with coauthoring, automatic versioning, Git integrations and RBAC.
- ✔ 12x better price/performance than cloud data warehouses
See why over 7,000 customers worldwide rely on Databricks for all their workloads from BI to AI.

Choose a cloud provider

Amazon Web Services

Microsoft Azure

Google Cloud Platform

Get started

By clicking "Get started" you agree to the [Privacy Policy](#) and [Terms of Service](#)

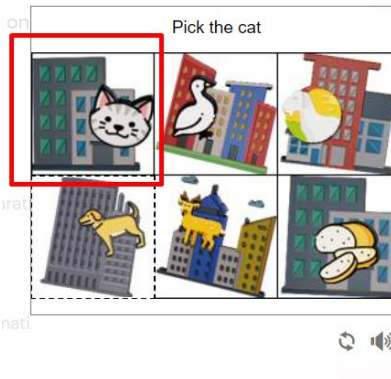
Don't have a cloud account?

Community Edition is a limited Databricks environment for personal use and training.

[Get started with Community Edition](#) →

By clicking "Get started with Community Edition" you agree to the [Privacy Policy](#) and [Terms of Service](#)

get your FOREVER FREE account by selecting this instead!!



now go to your email!

Check your email to start your trial.

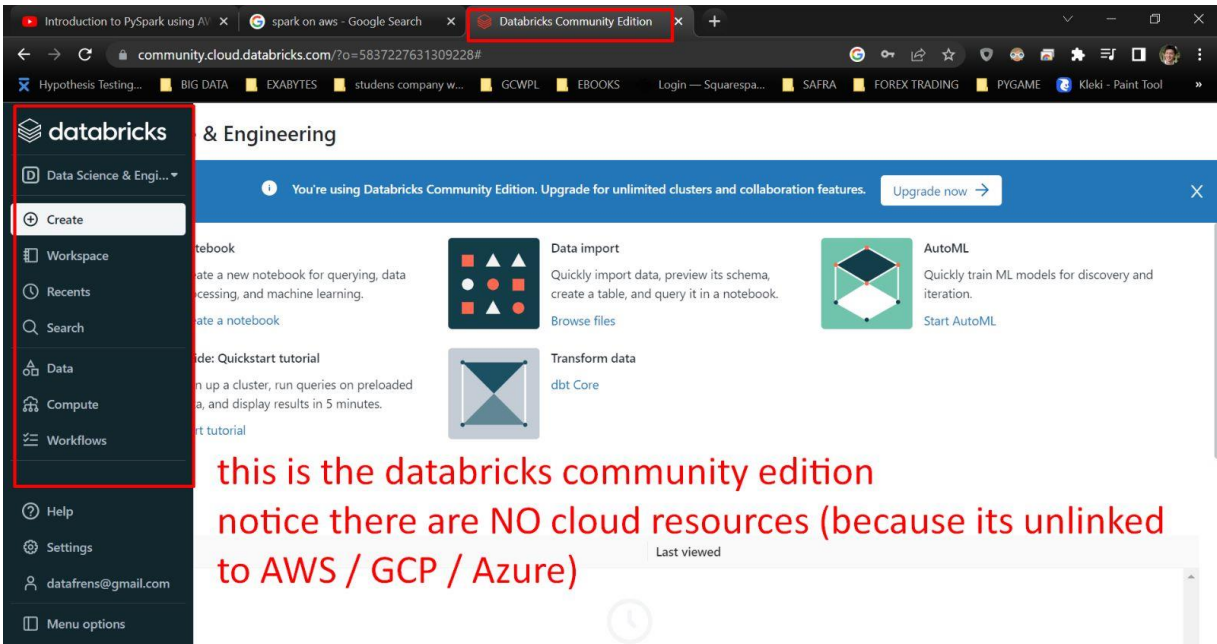
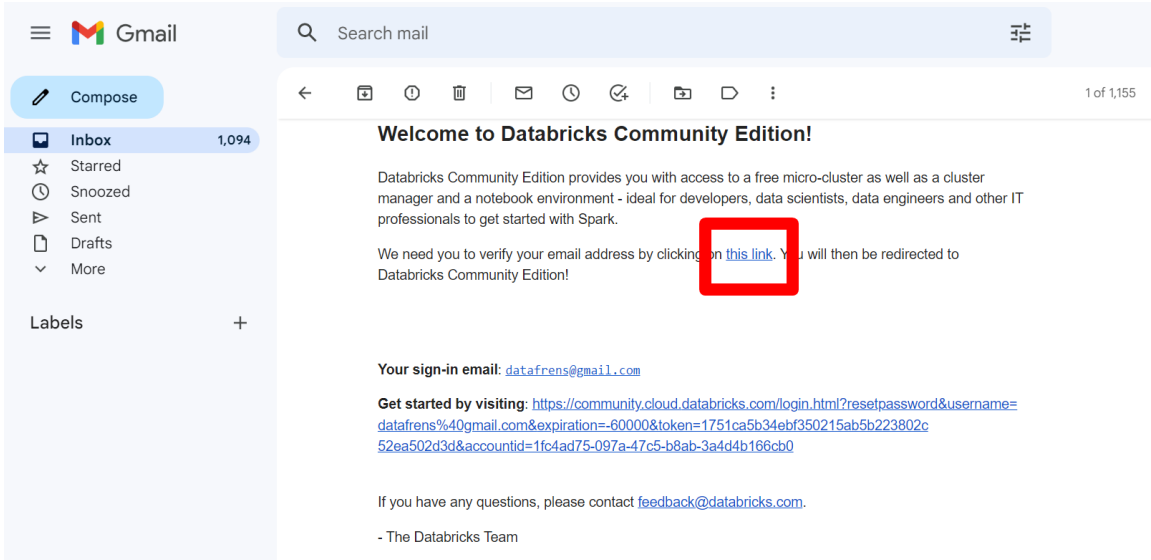
Thank you for signing up. Please validate your email address to start your trial.

Here are some resources to help you deploy your first workspace.

1. [Review the administration guide](#) on the requirements to set up your Databricks service.
 - Not an admin on your AWS Account? Share [this guide](#) with your admin to deploy a workspace for you!
2. [Follow our Quickstart guide to create your first workspace.](#)

You can also check out our [Docs](#) and [Community](#) sites to get your questions answered.

Note: if you signed up for Community Edition, you'll go to your first workspace as soon as you verify your



ABOUT DR. ALVIN ANG



Dr. Alvin Ang earned his Ph.D., Masters and Bachelor degrees from NTU, Singapore. He is a scientist, entrepreneur, as well as a personal/business advisor. More about him at www.AlvinAng.sg.