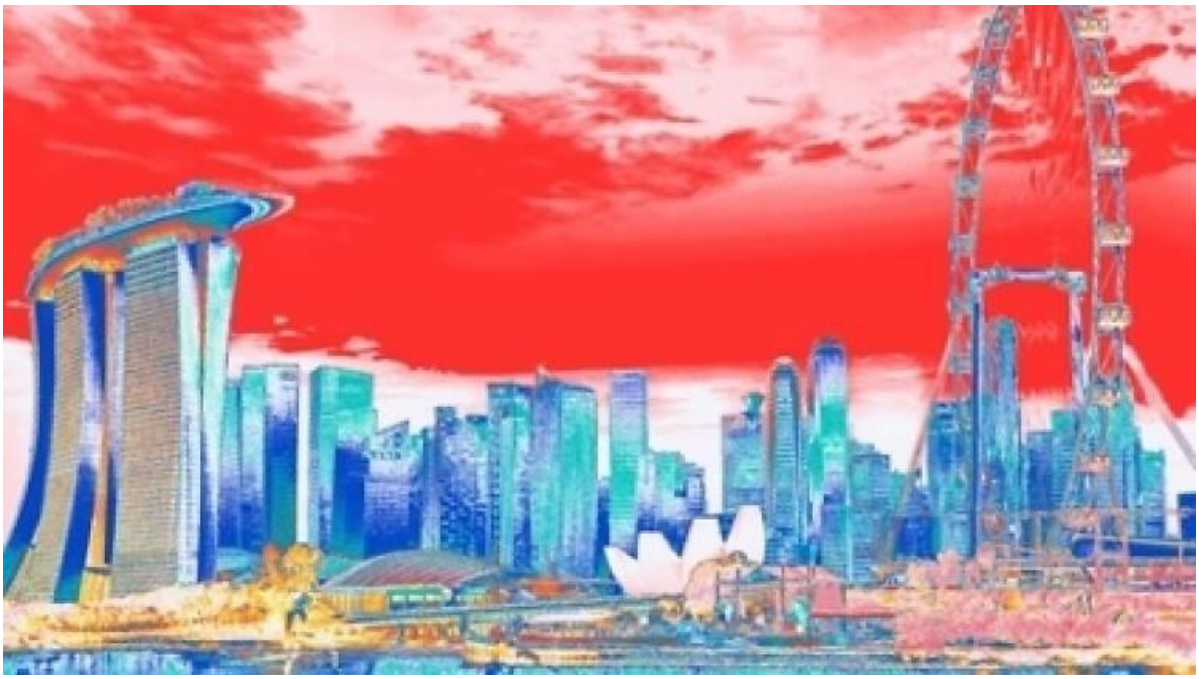
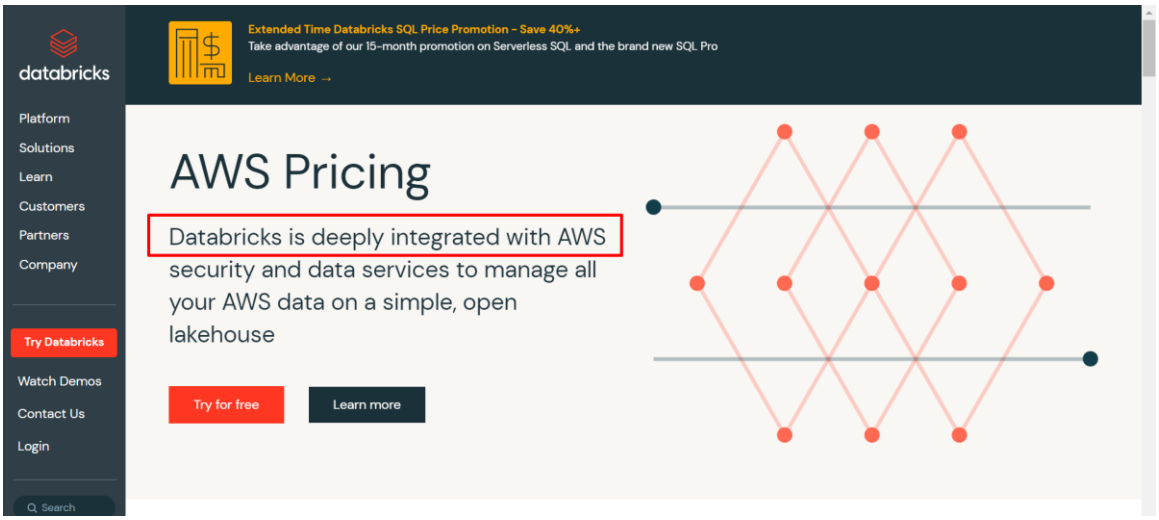
USING APACHE SPARK IN AWS AND DATABRICKS

DR. ALVIN ANG

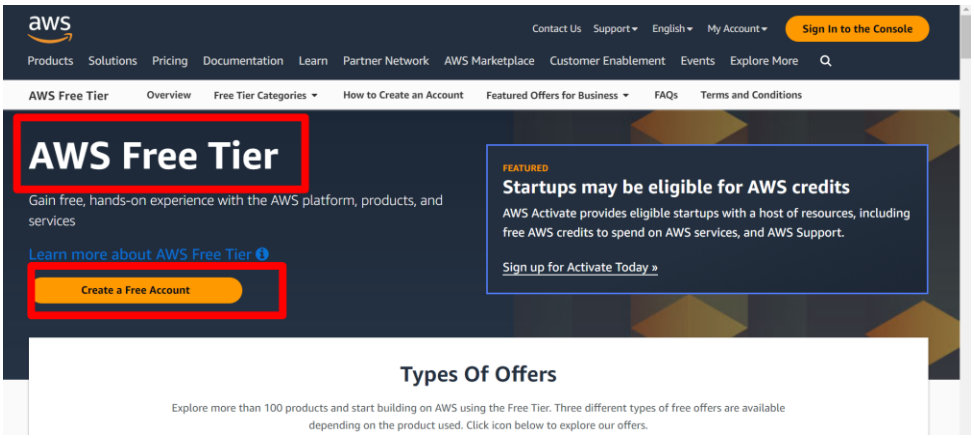


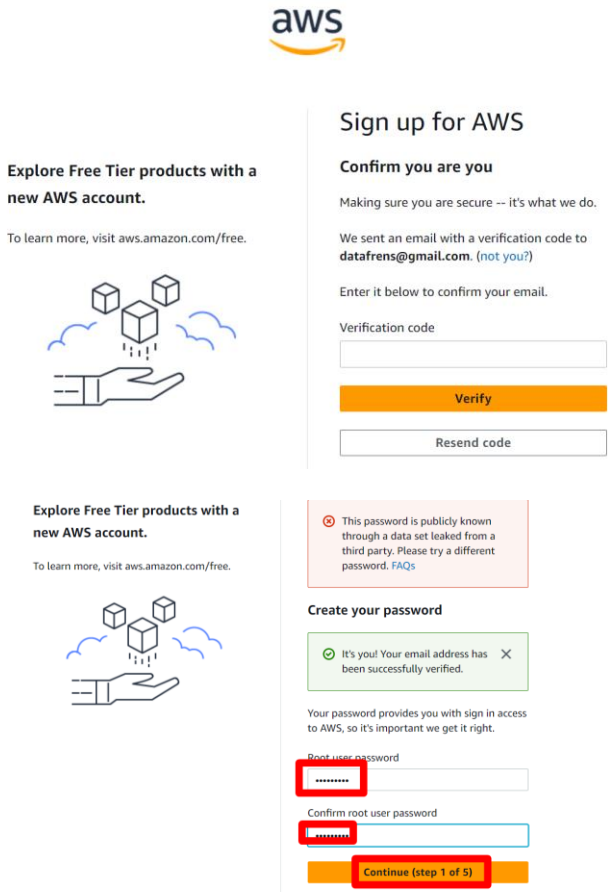
CONTENTS

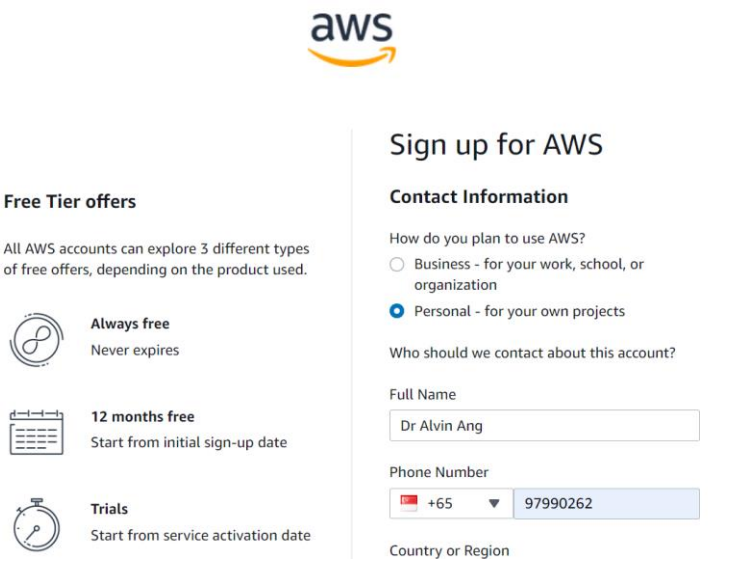


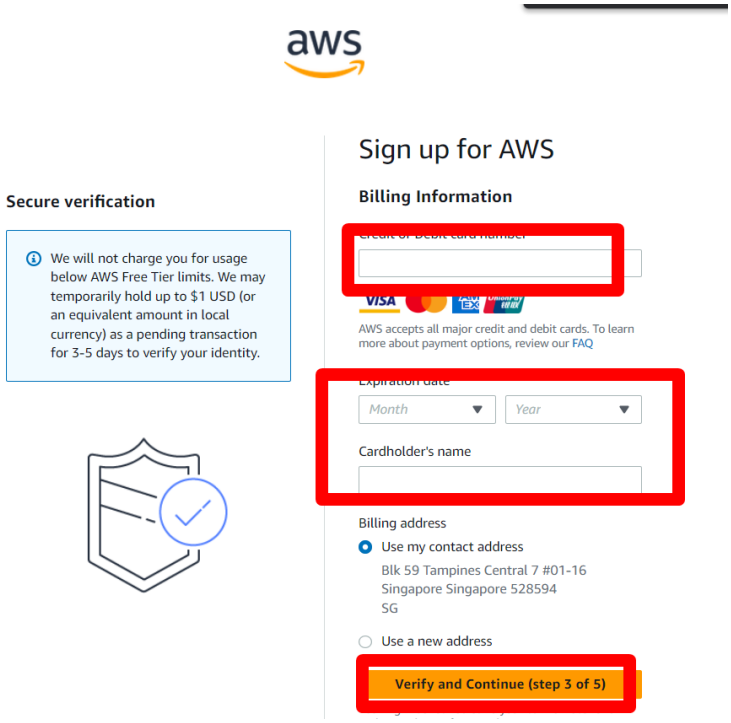
- Realistically, Spark needs to run on clusters to do parallel computing.
- Databricks is an easy way of managing the complex Spark architecture.
- We shall use Databricks as the Jupyter Notebook interface [frontend] to run the codes and manage the clusters, but use AWS to form those clusters and store 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.

ign Up with AWS

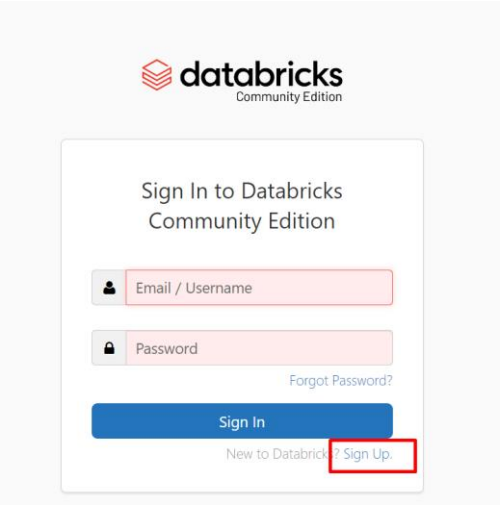




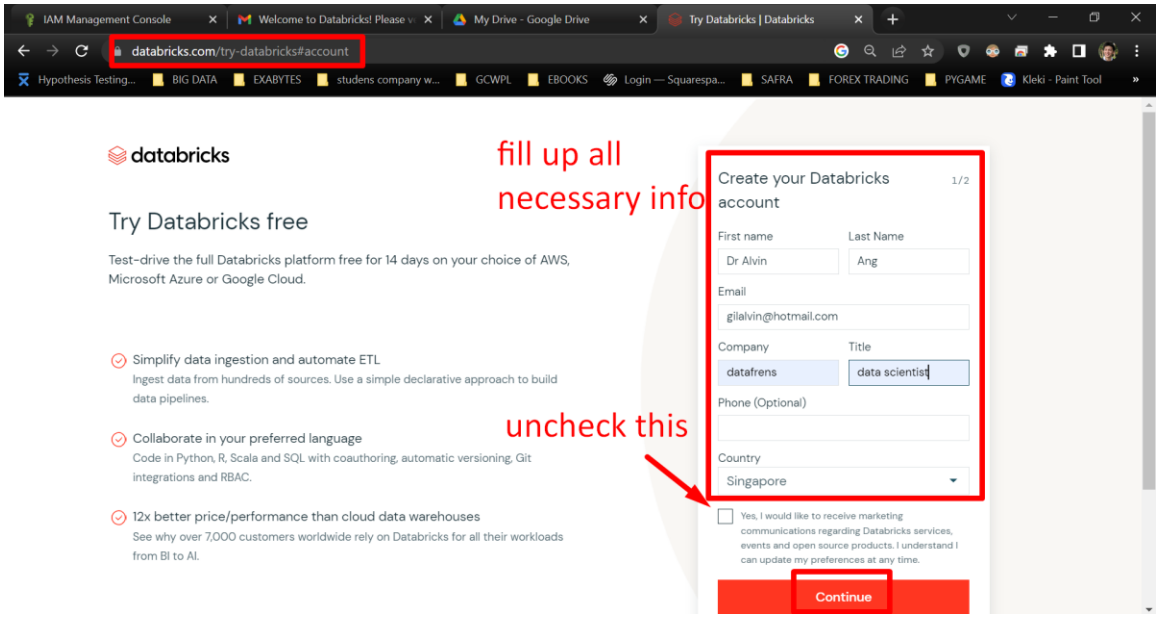


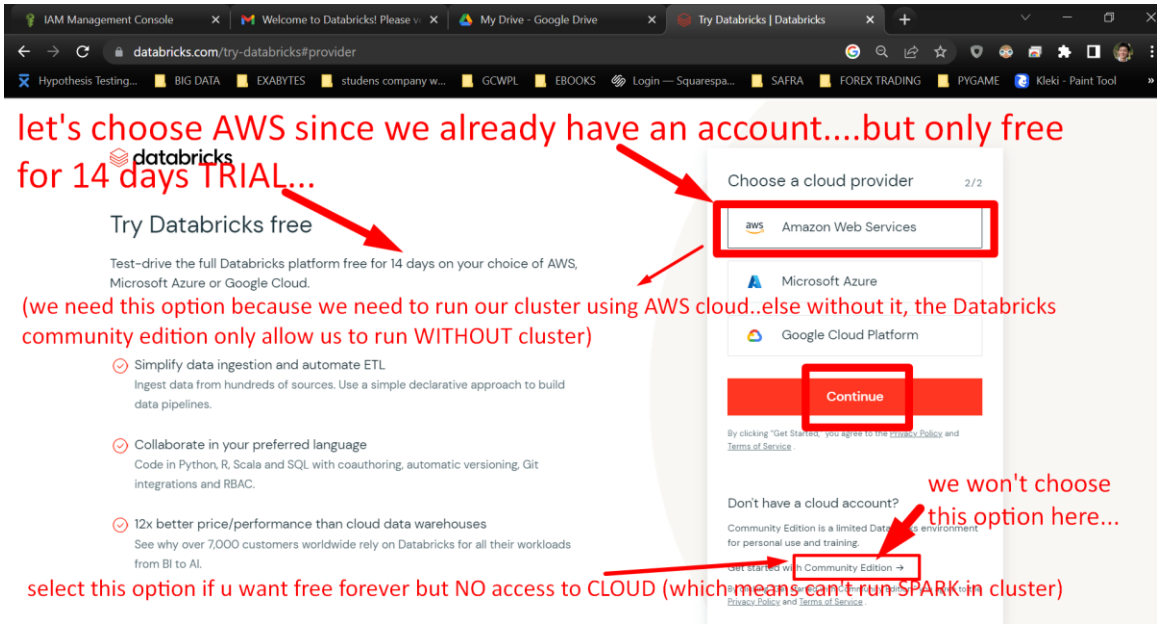


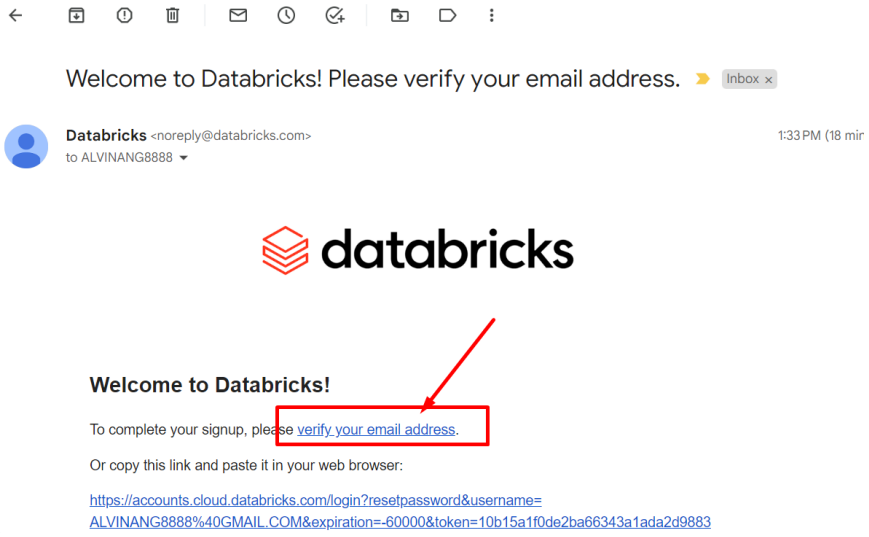
II. SIGN UP WITH DATABRICKS (14 DAYS FREE TRIAL FULL ACCOUNT)

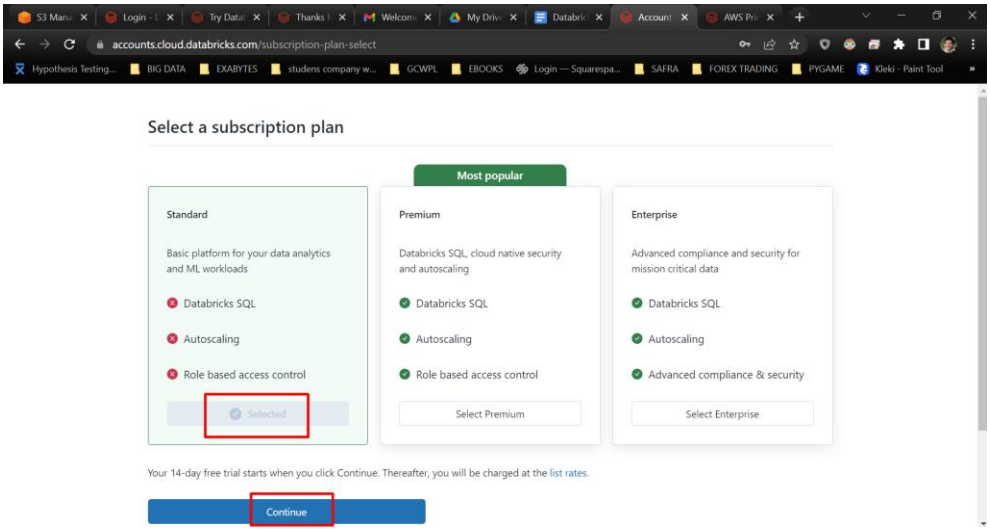


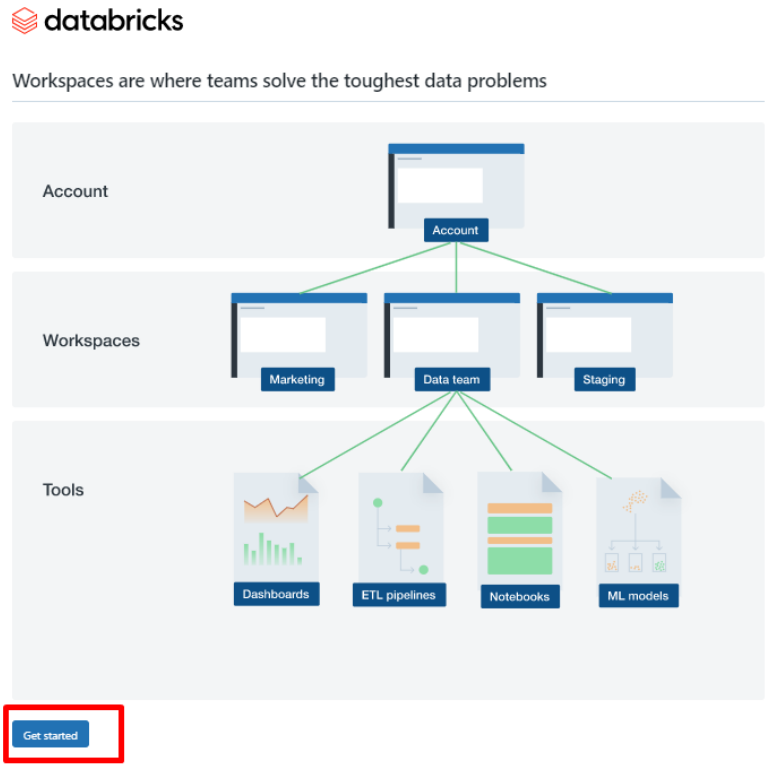
https://www.databricks.com/try-databricks#account

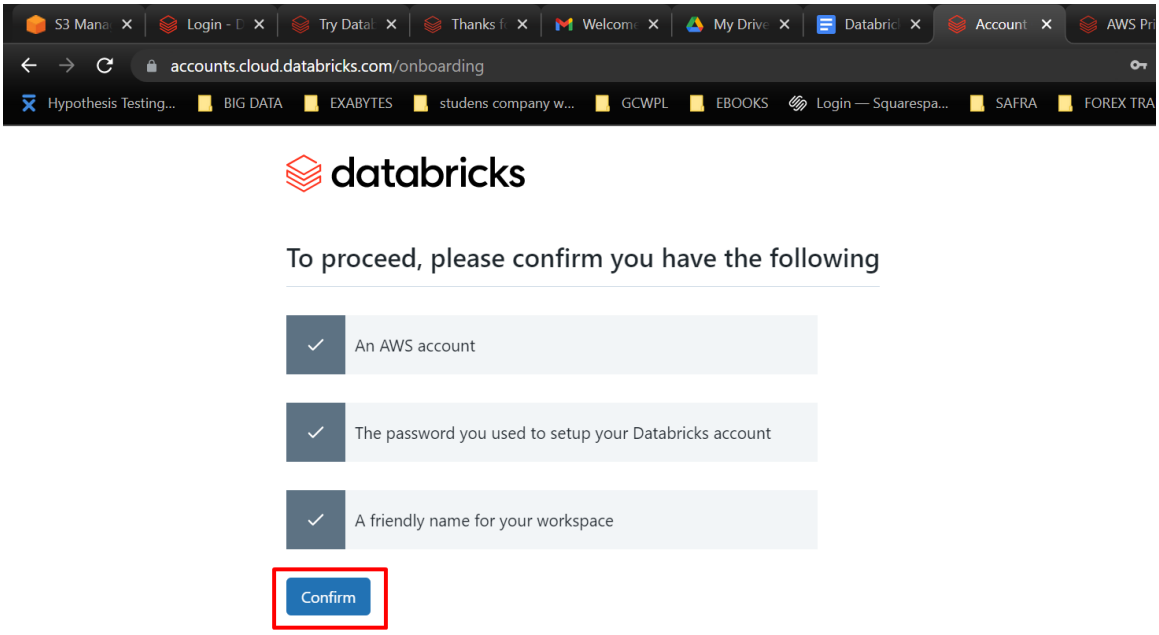


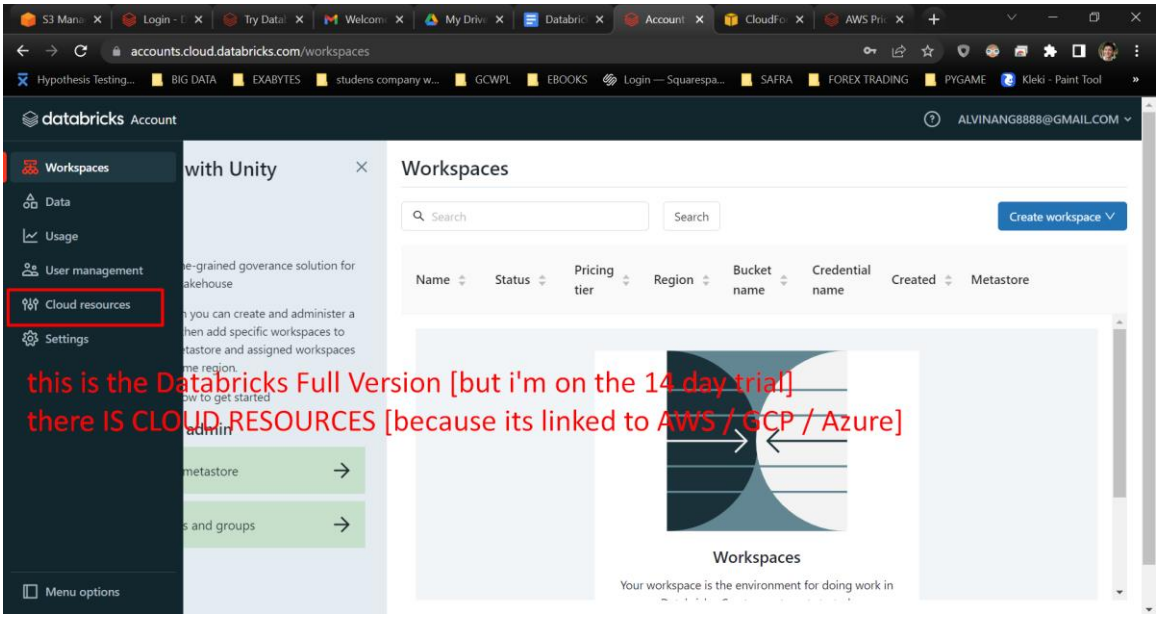








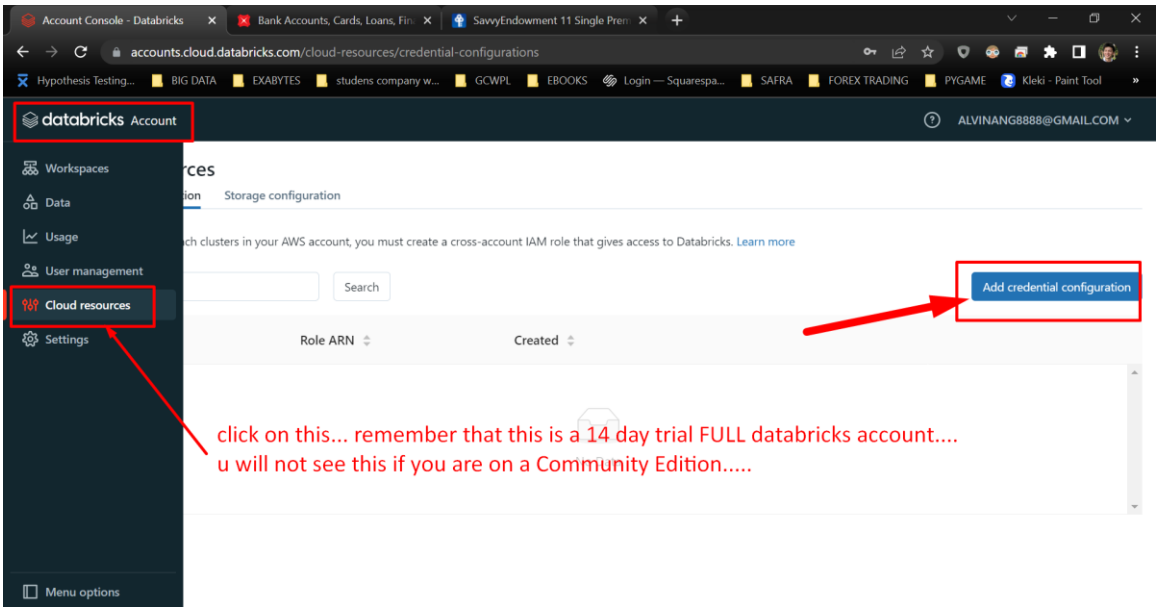


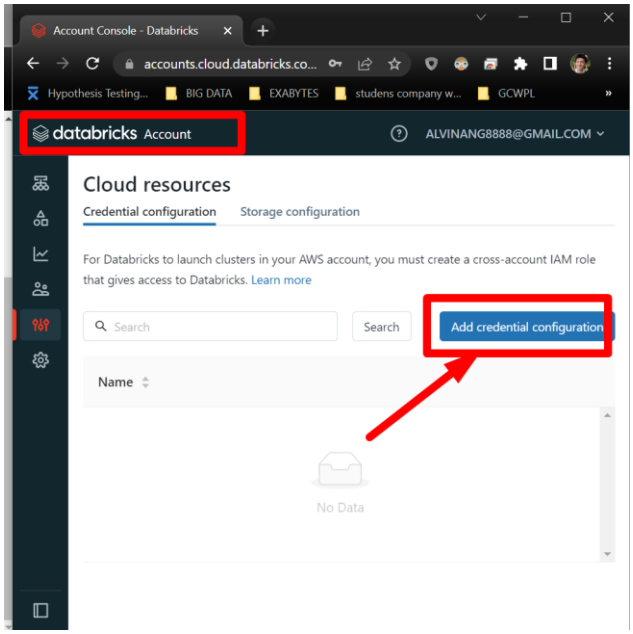


III. CREATING THE CROSS ACCOUNT IAM ROLE

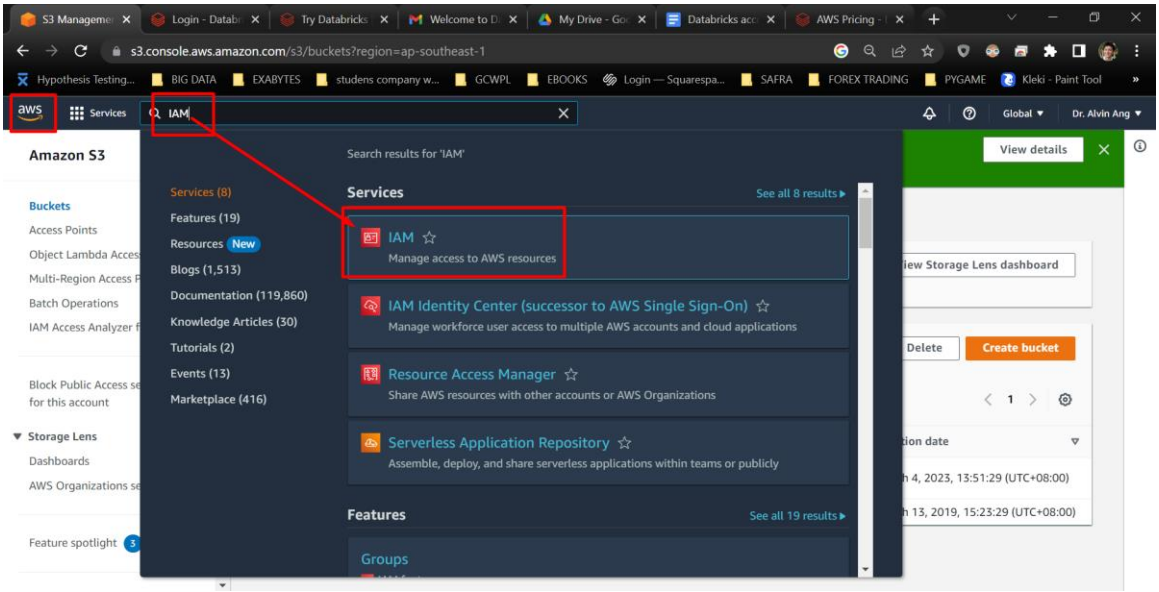
https://docs.databricks.com/administration-guide/account-api/iam-role.html#language-Databricks%C2%A0VPC

A. ADD CREDENTIAL CONFIGURATION FROM DATABRICKS

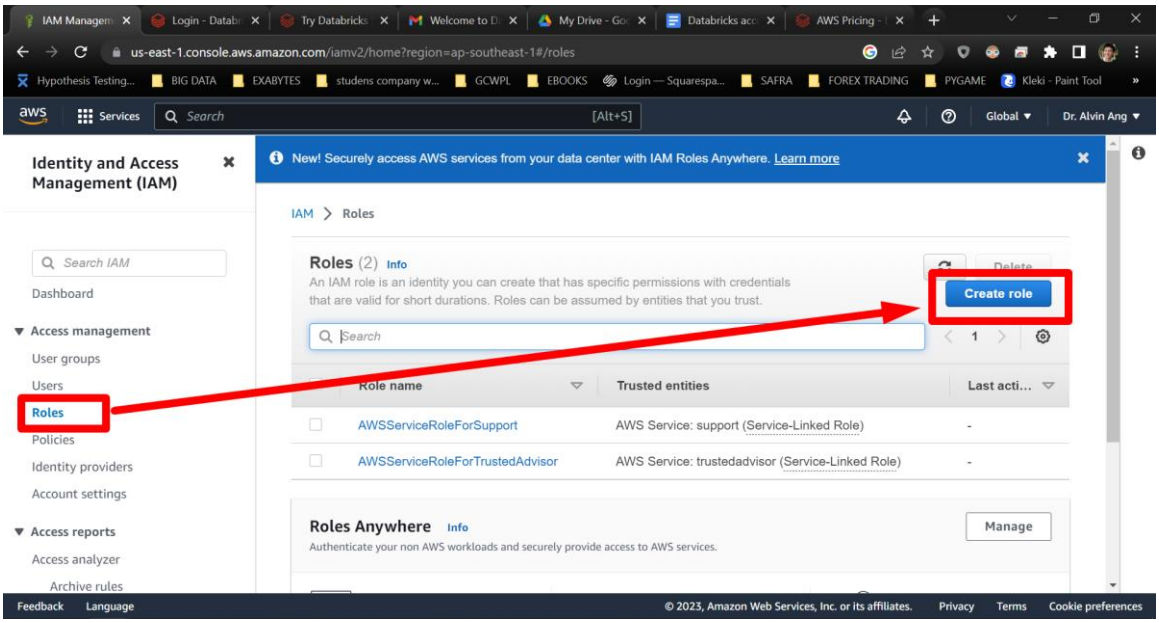




B. GO TO IAM IN AWS

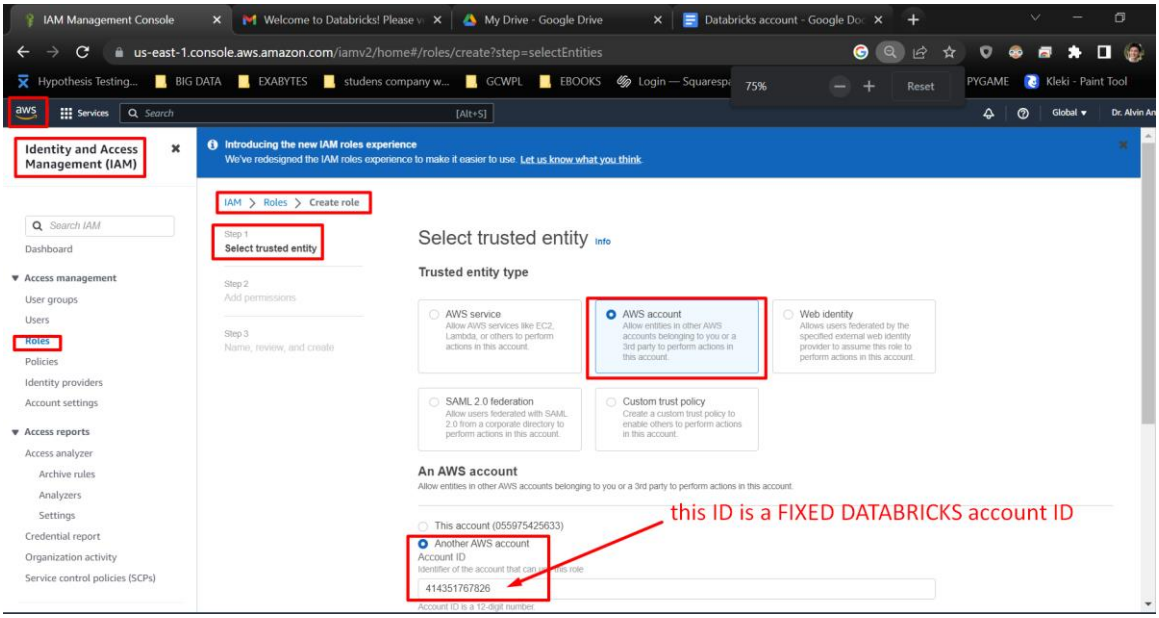


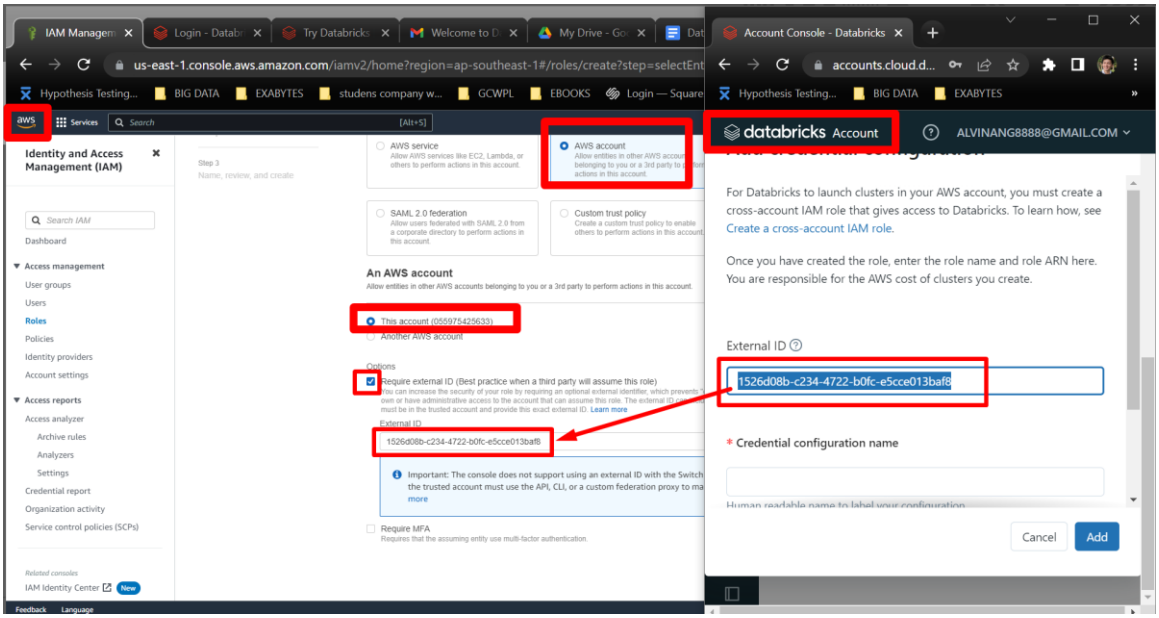
C. CREATE ROLE IN AWS

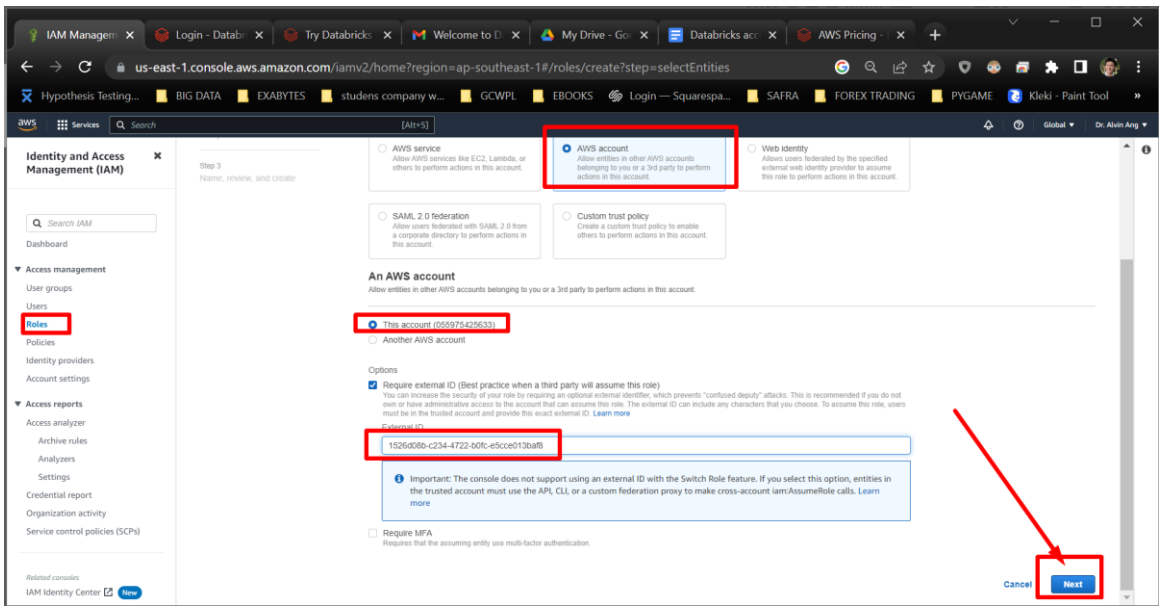


D. ENTER THE DATABRICKS ACCOUNT ID IN AWS

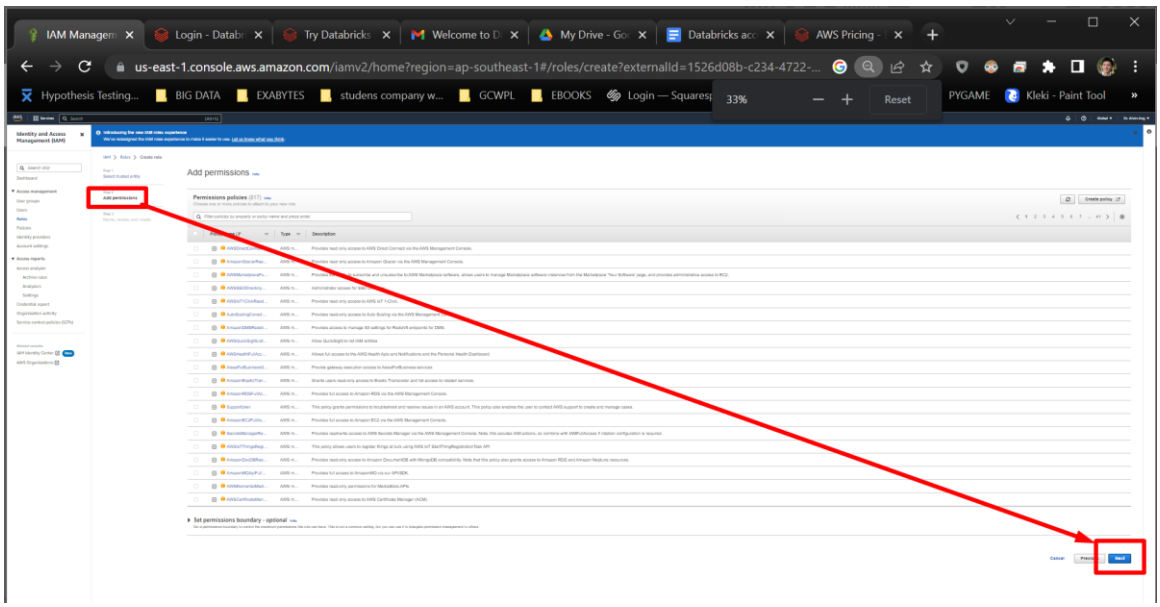
Databricks account ID = 414351767826



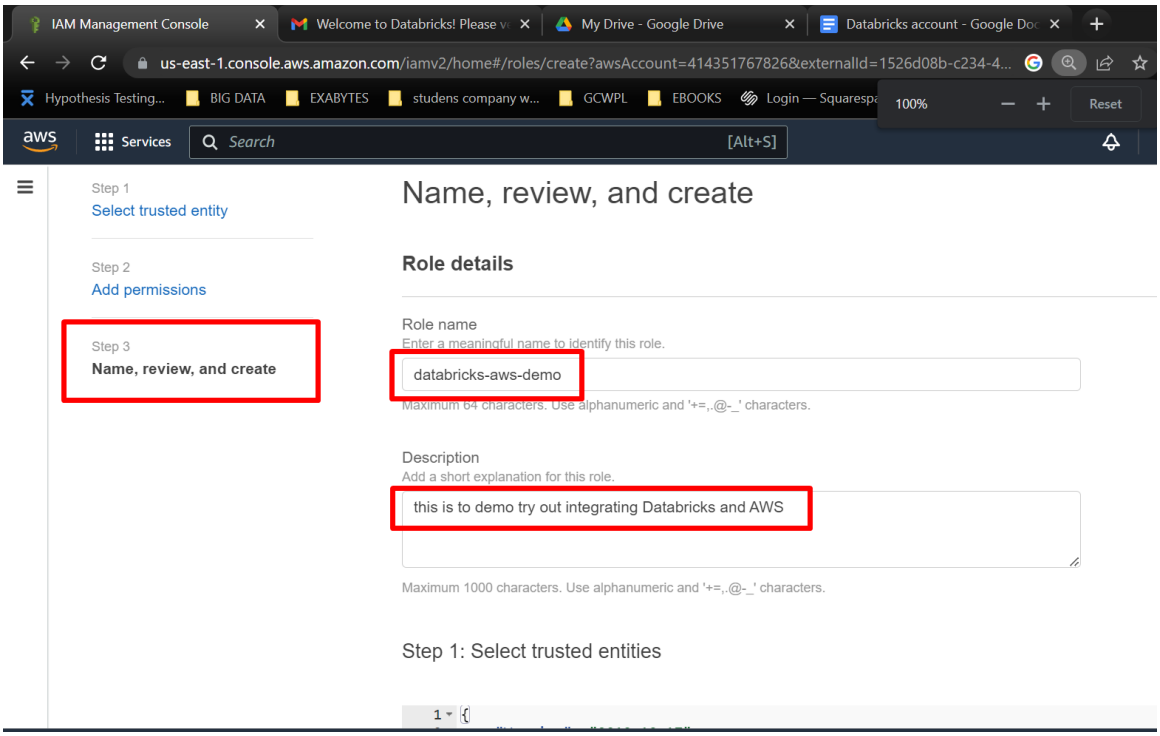


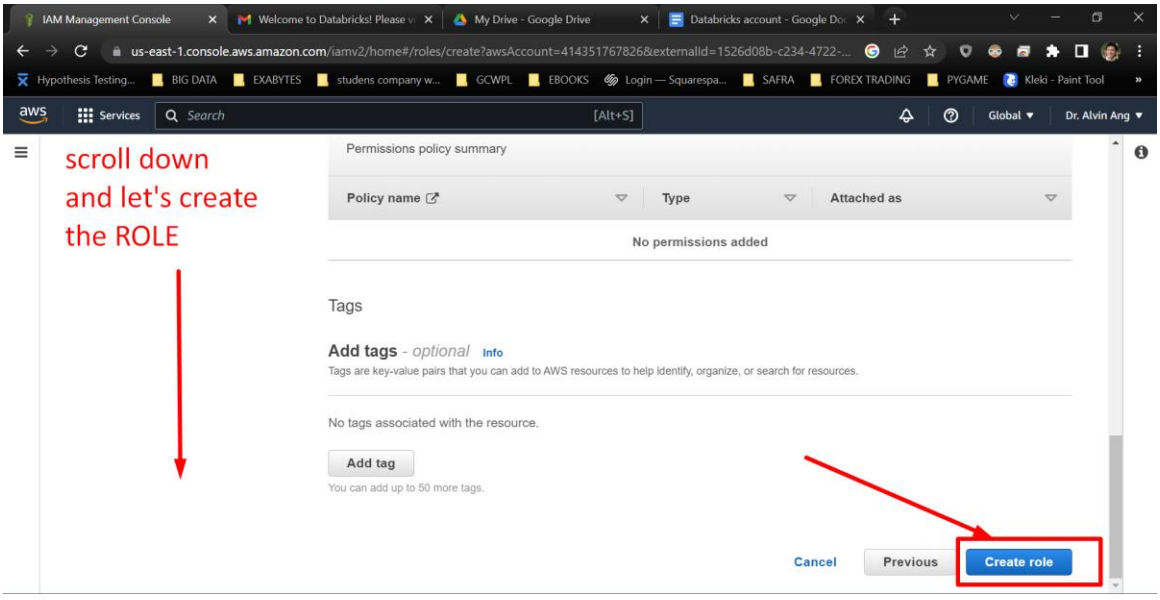


E. IGNORE THE STEP 2: ADD PERMISSIONS AND JUST CLICK NEXT

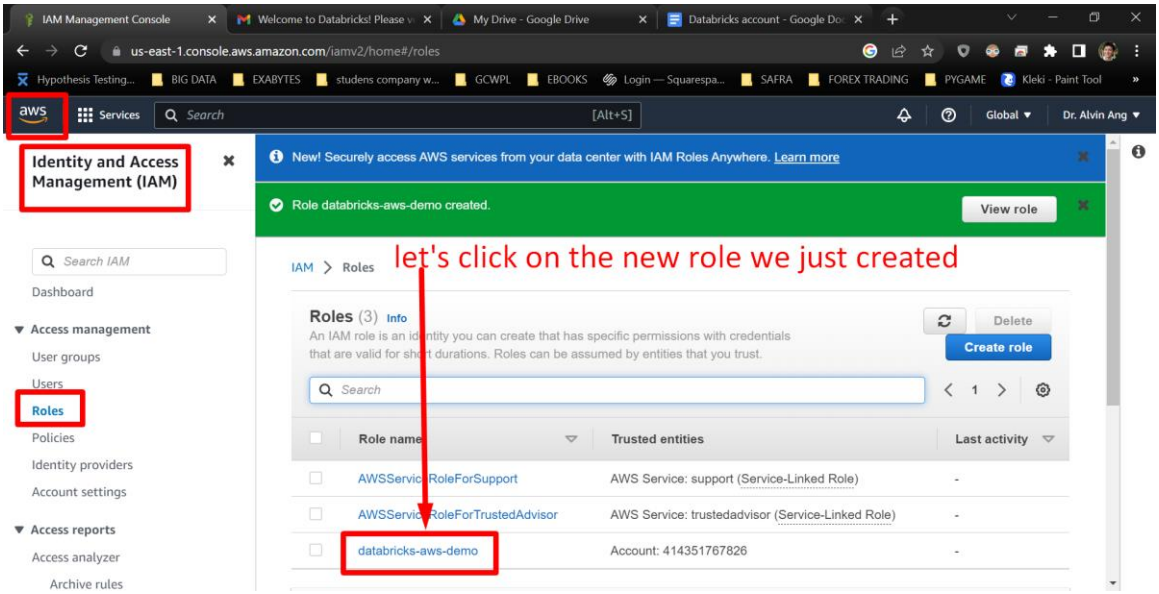


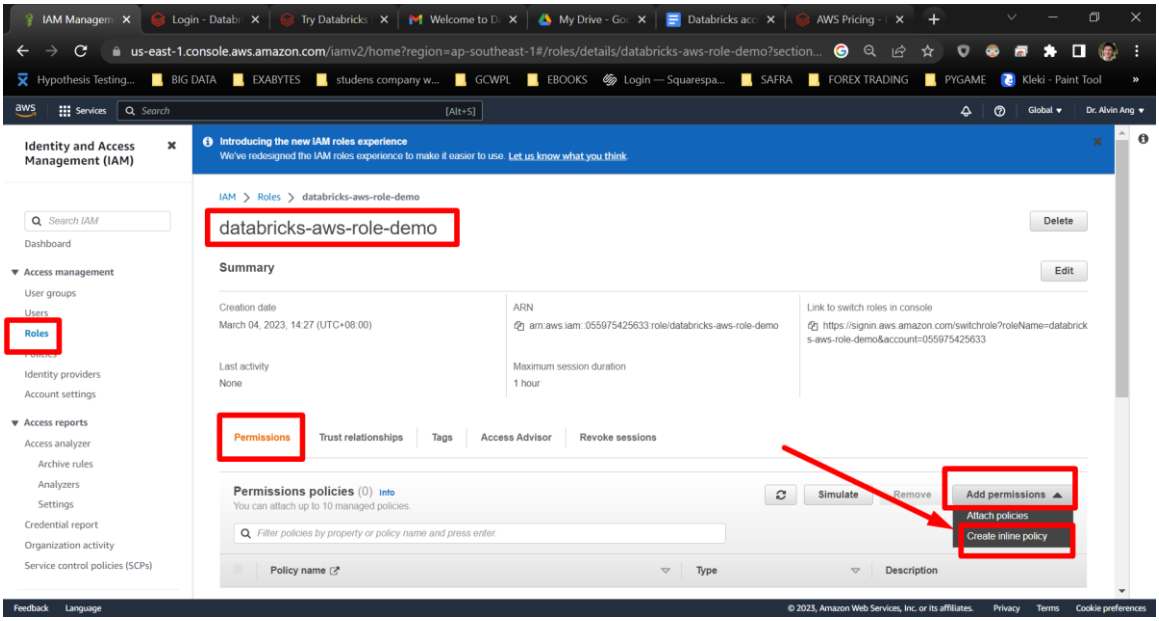
F. ENTER A ROLE NAME AND DESCRIPTION



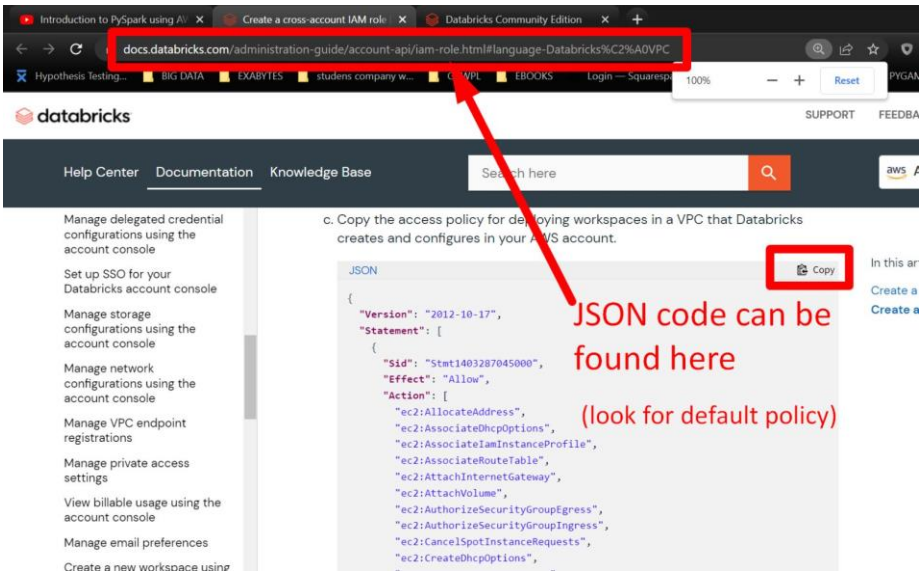


G. CREATE INLINE POLICY IN AWS NEW ROLE



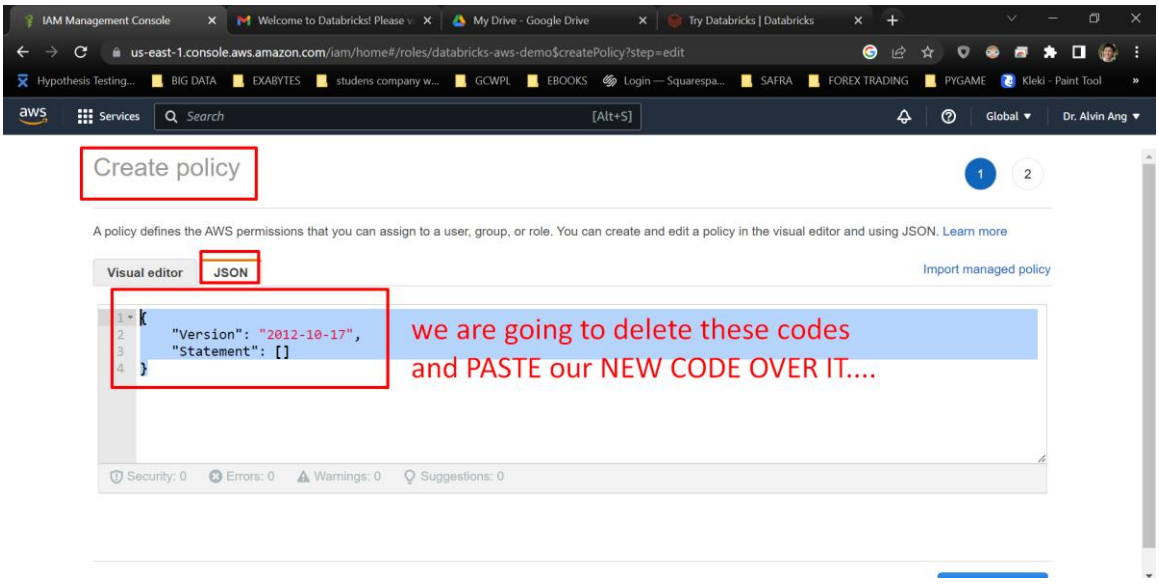


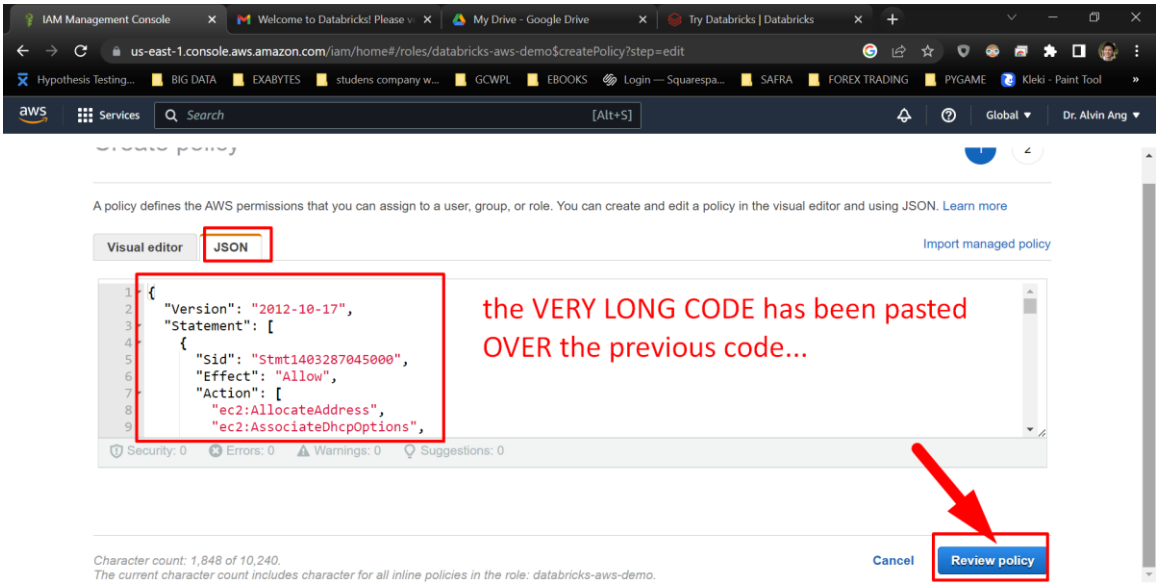
H. COPY JSON CODE



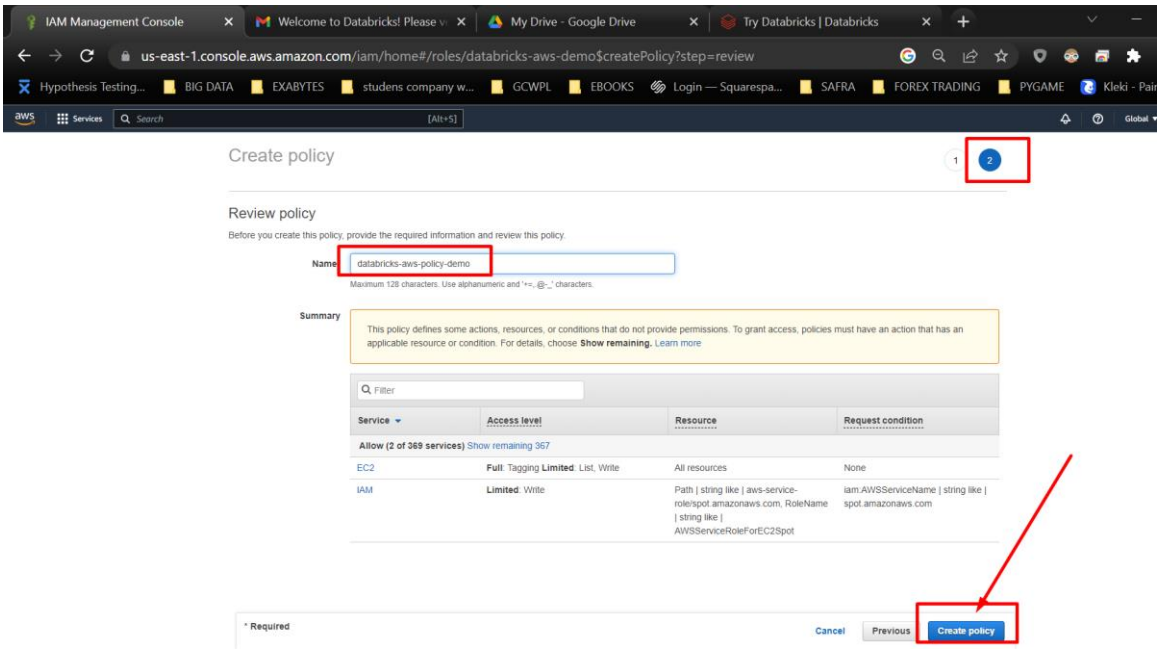
https://docs.databricks.com/administration-guide/account-api/iam-role.html#language-Databricks%C2%A0VPC

I. PASTE JSON CODE

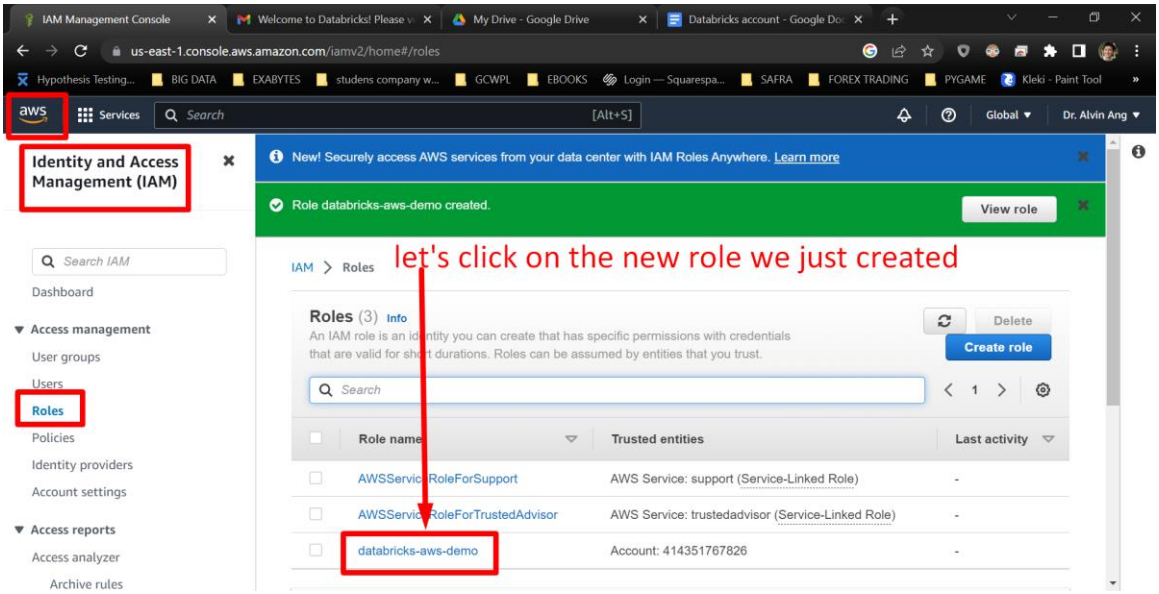


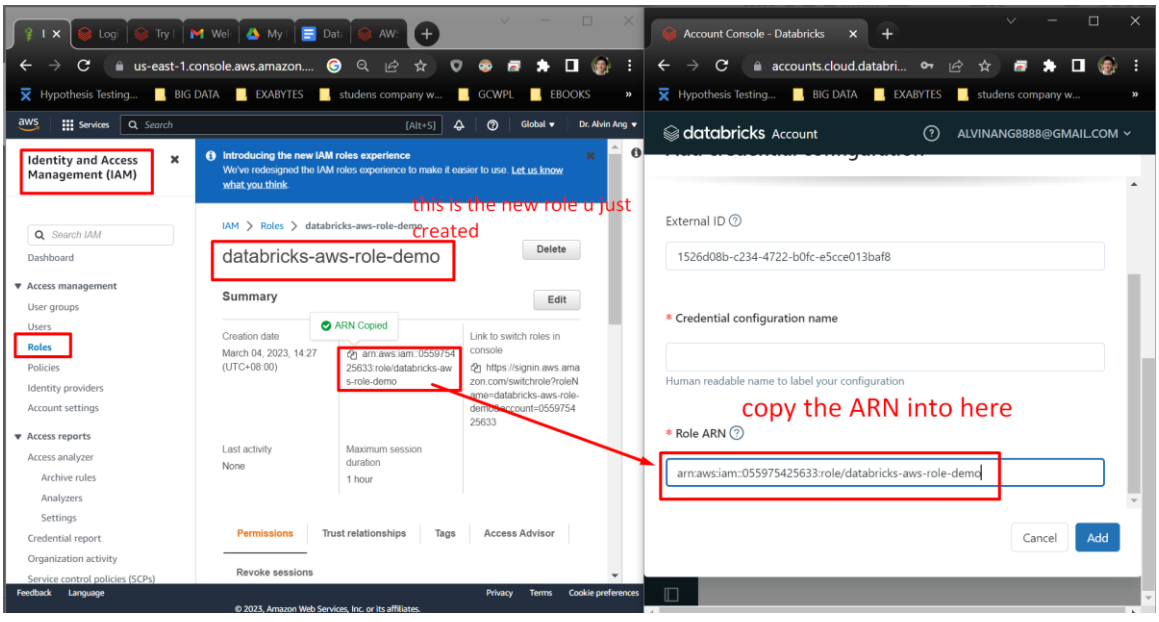


```
{
 "Version": "2012-10-17",
 "Statement": [
  {
   "Sid": "Stmt1403287045000",
   "Effect": "Allow",
   "Action": [
    "ec2:AllocateAddress",
    "ec2:AssociateDhcpOptions",
    "ec2:AssociateIamInstanceProfile",
    "ec2:AssociateRouteTable",
    "ec2:AttachInternetGateway",
    "ec2:AttachVolume",
    "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"
    }
   }
  }
 ]
}
```

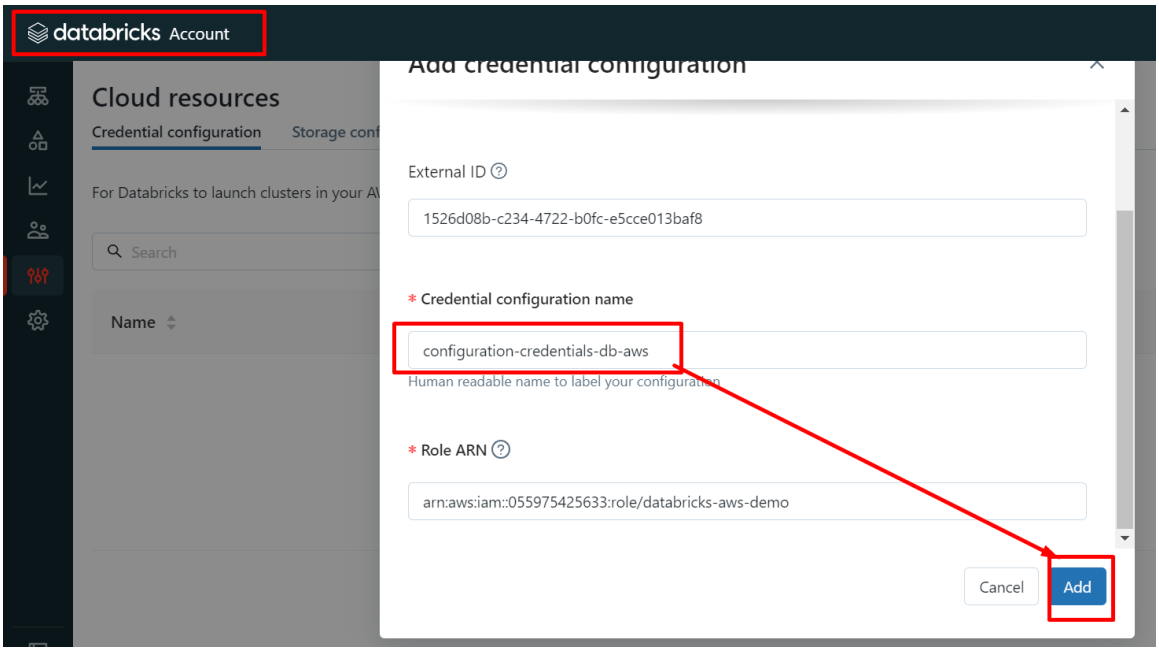


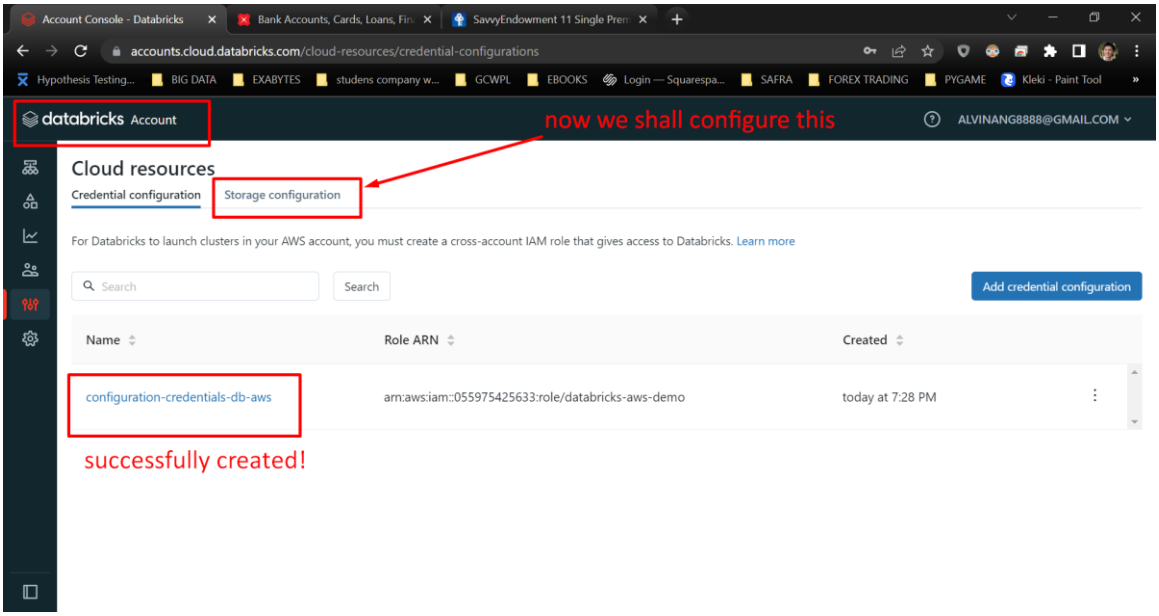
J. COPY OUT THE AMAZON RESOURCE NAME [ARN] FROM THE NEW ROLE (FROM AWS) INTO DATABRICKS

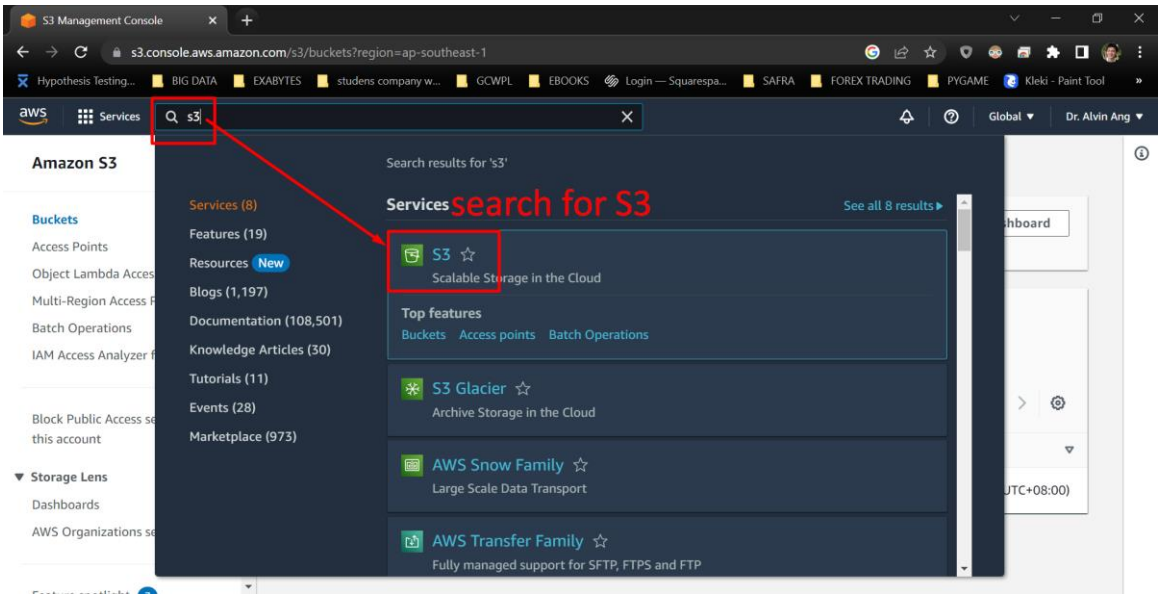


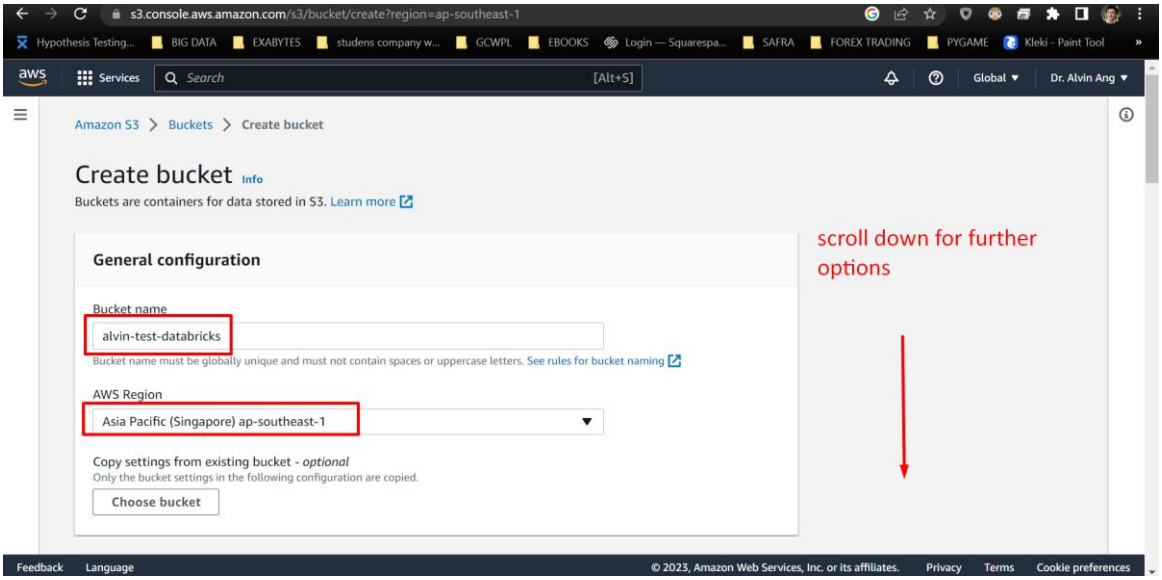


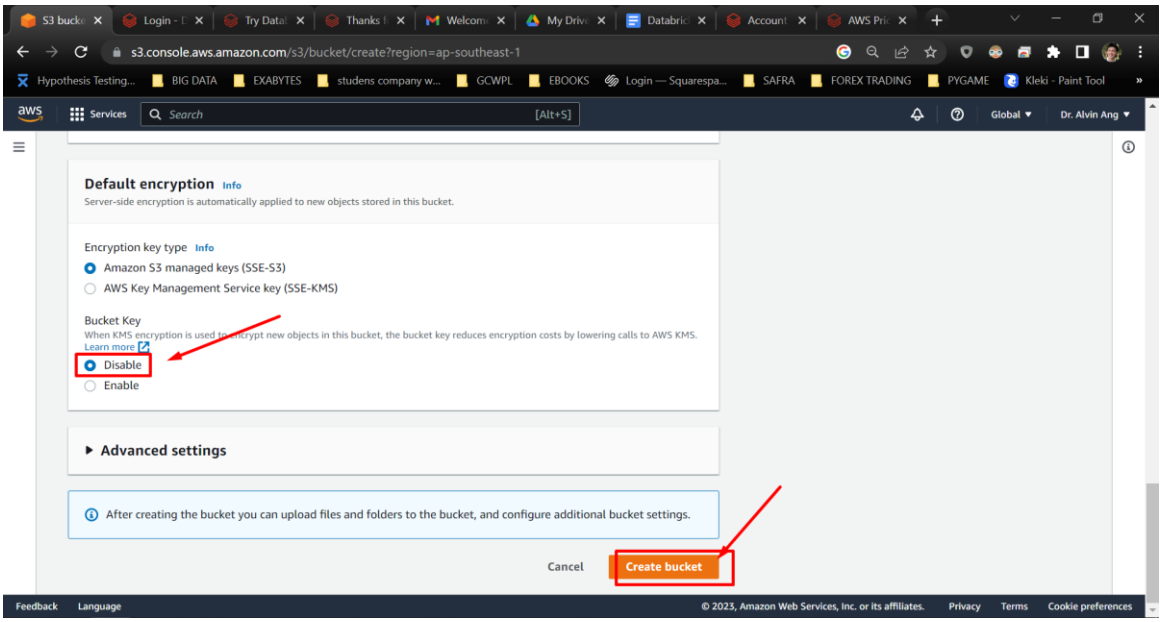
K. CREATE A CONFIGURATION NAME

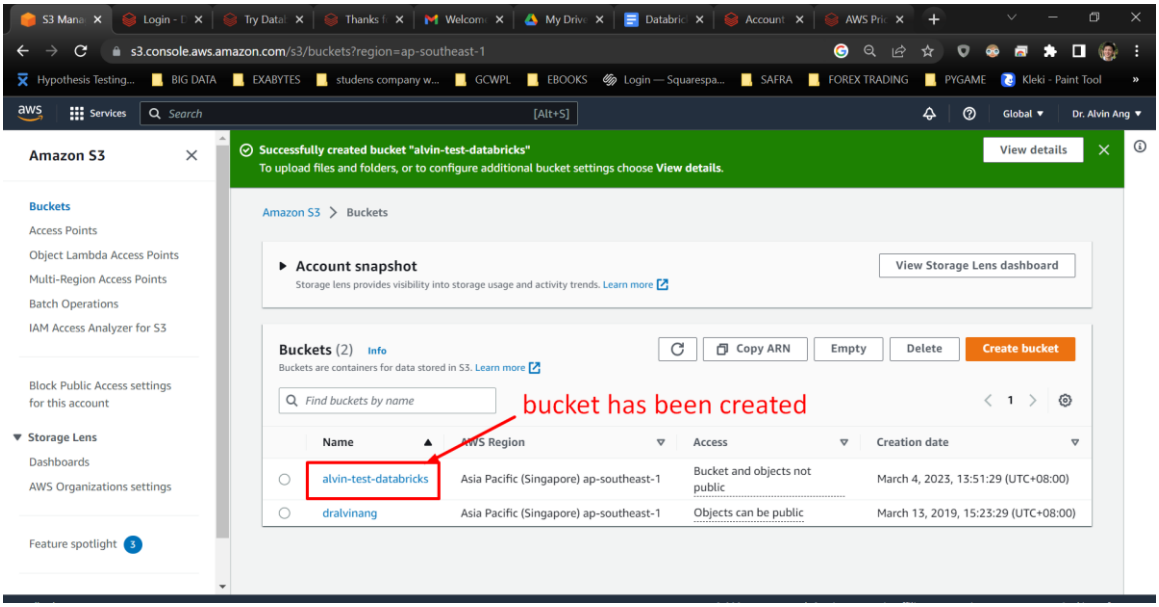








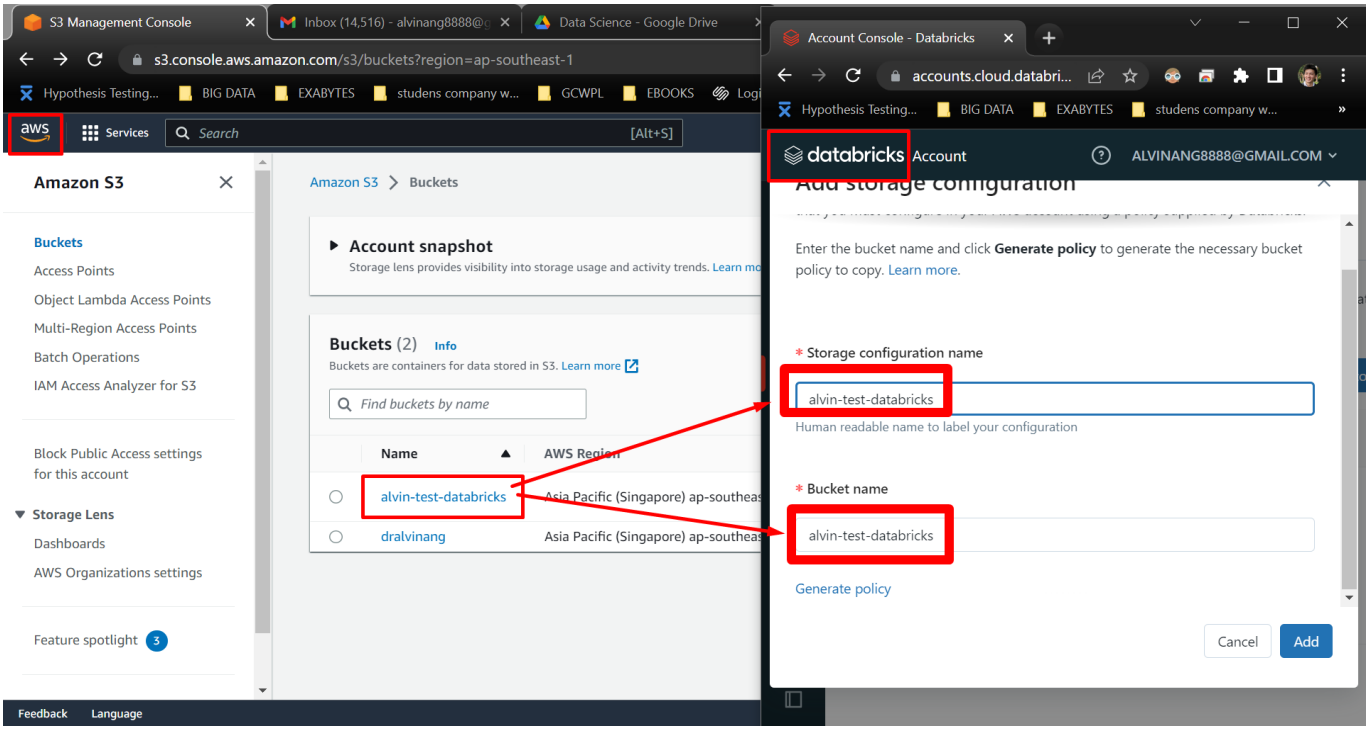




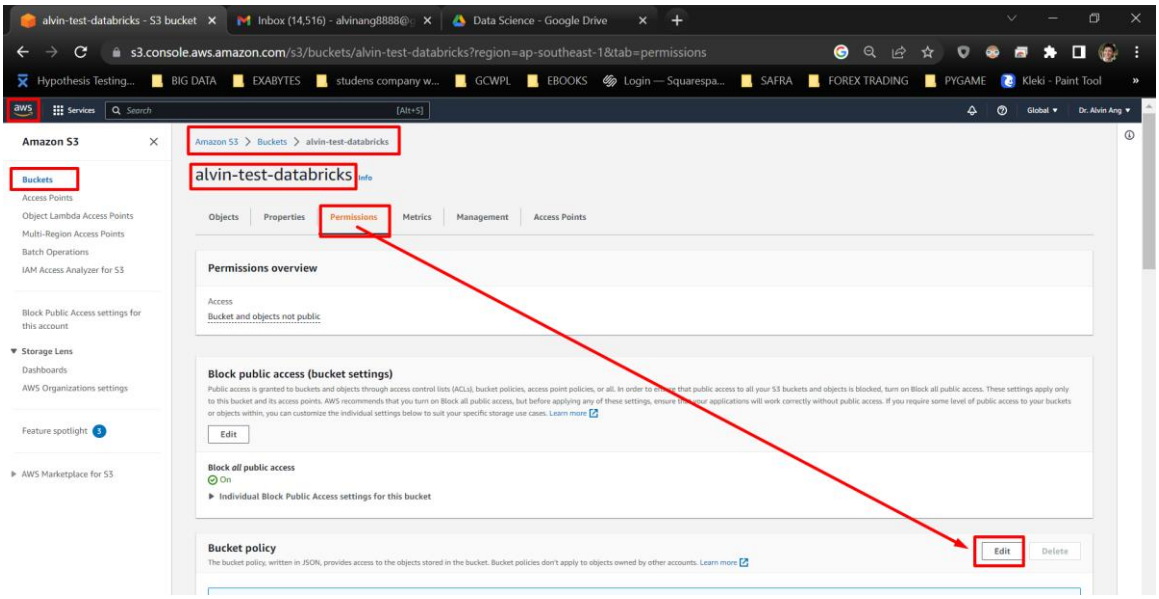
V. DATABRICKS STORAGE CONFIGURATION

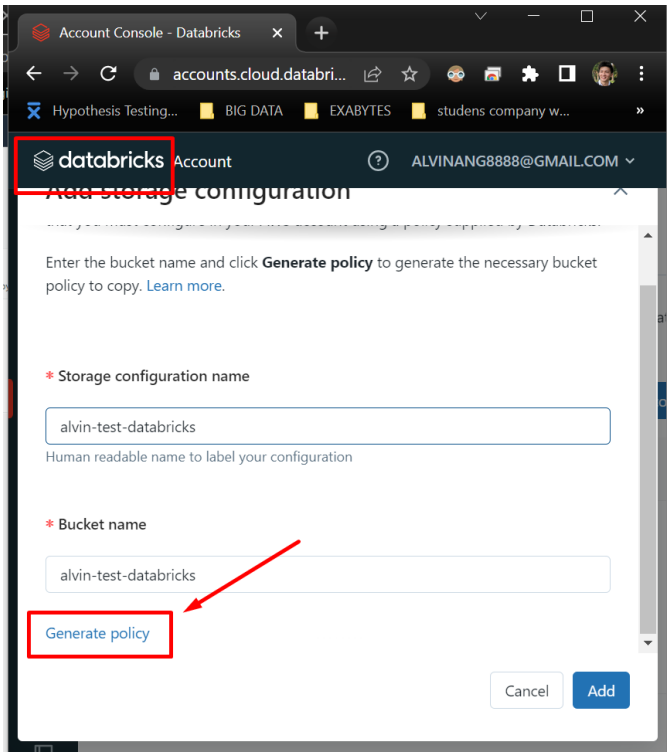
A. STORAGE CONFIGURATION NAME AND BUCKET NAME

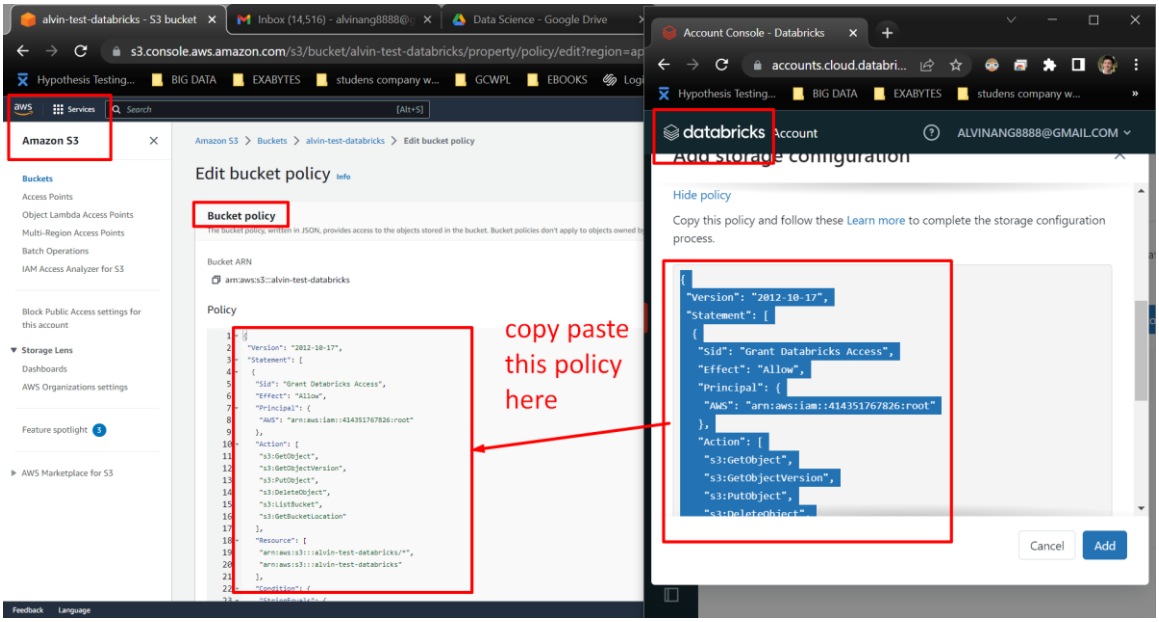
Recall previously the BUCKET name is called 'alvin-test-databricks'

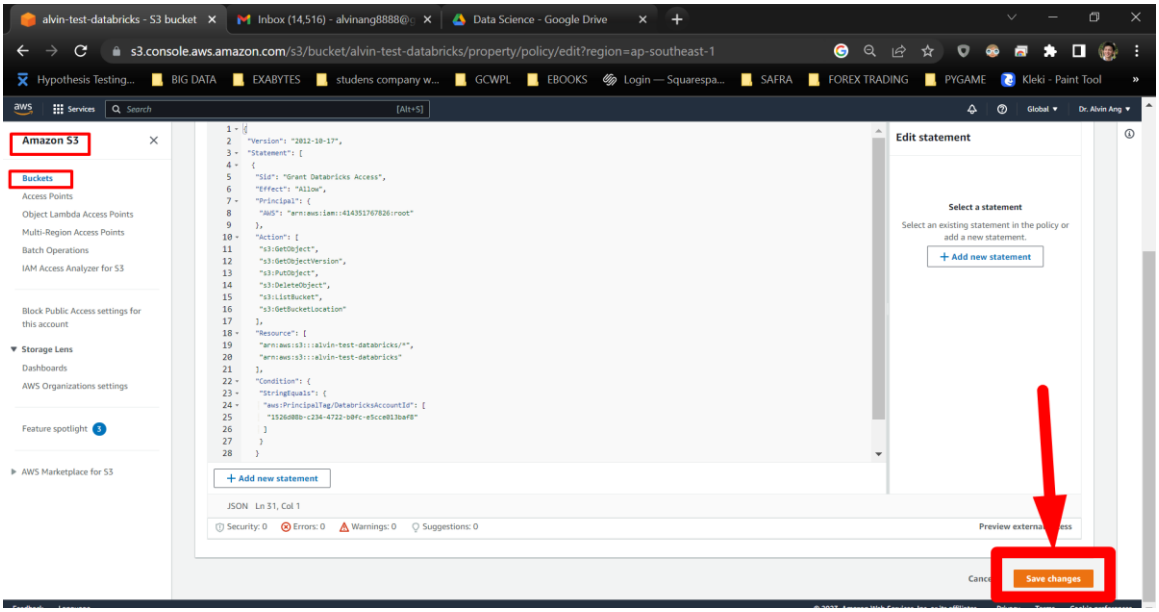


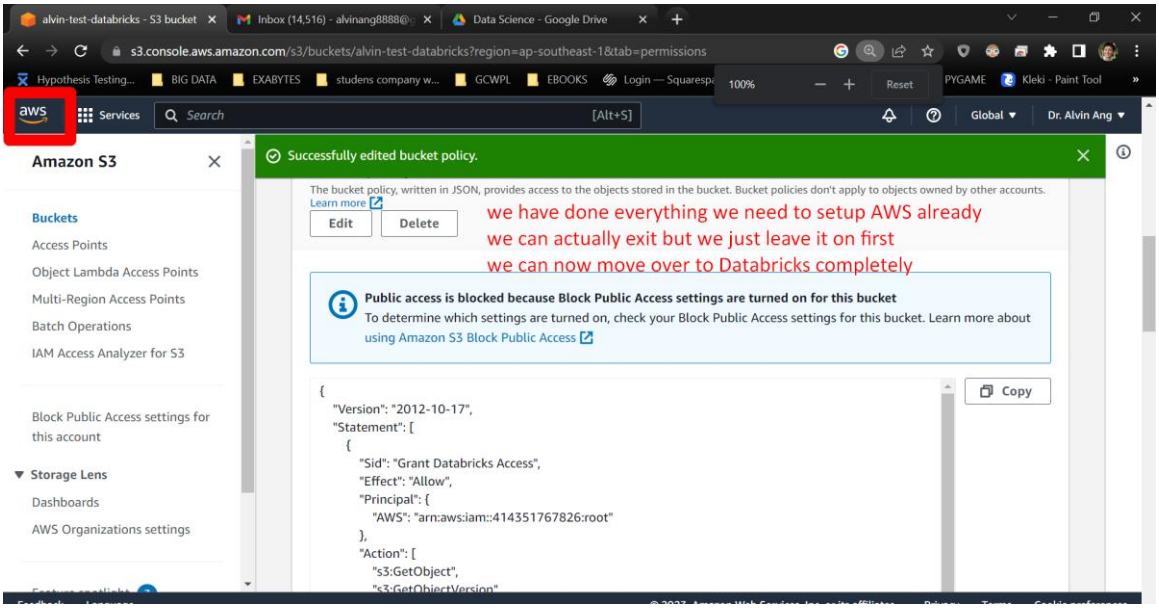
B. BUCKET POLICY



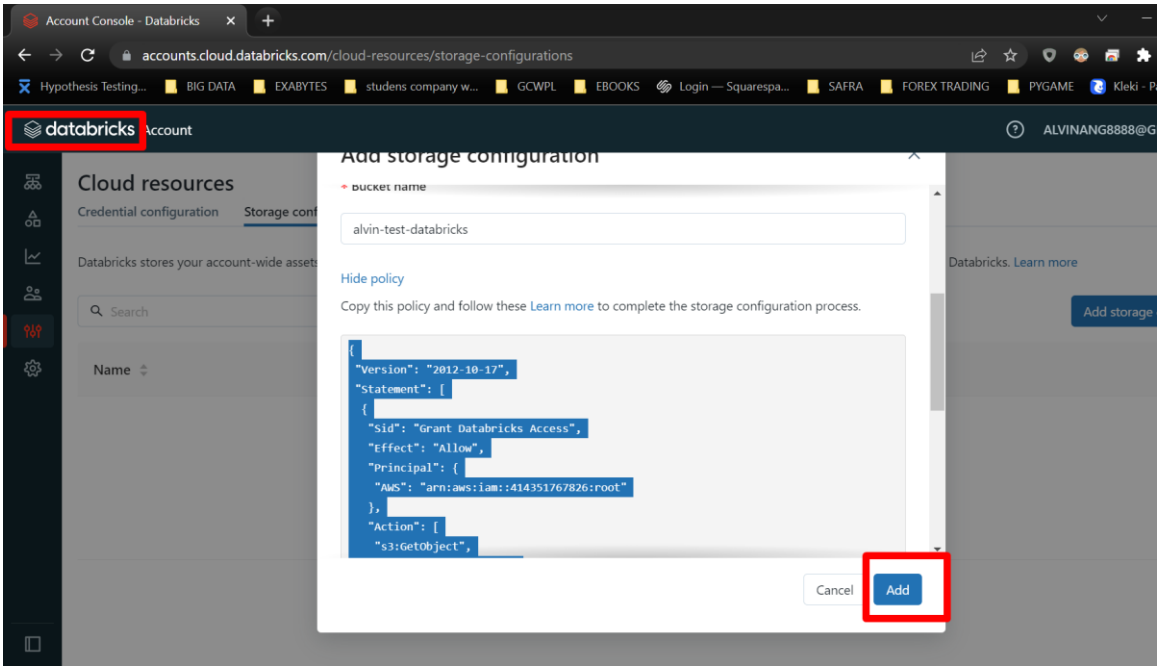


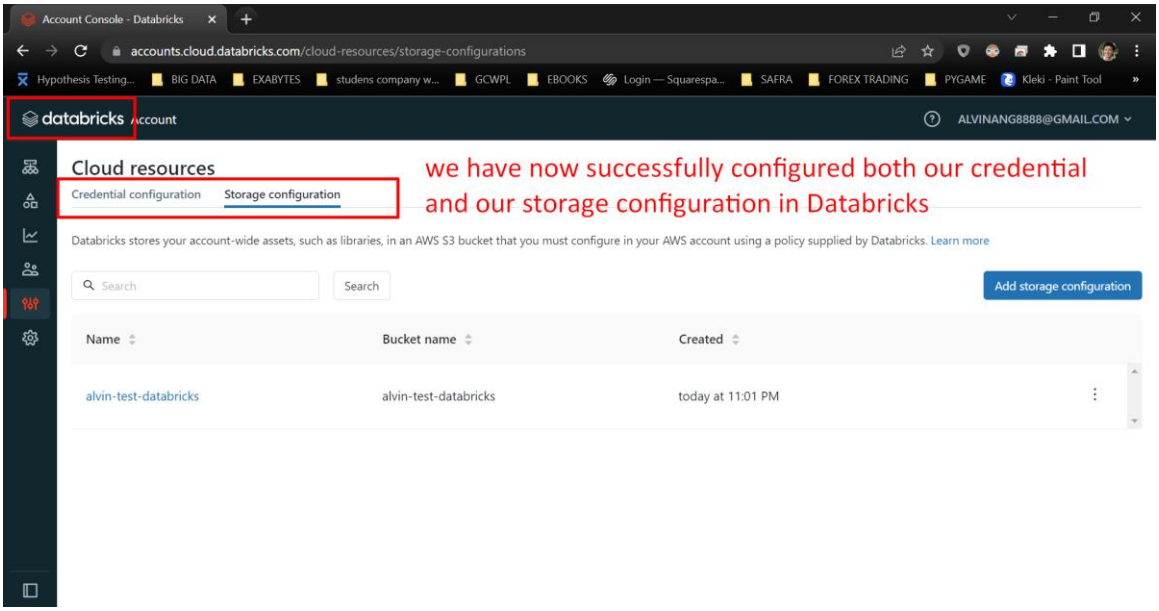






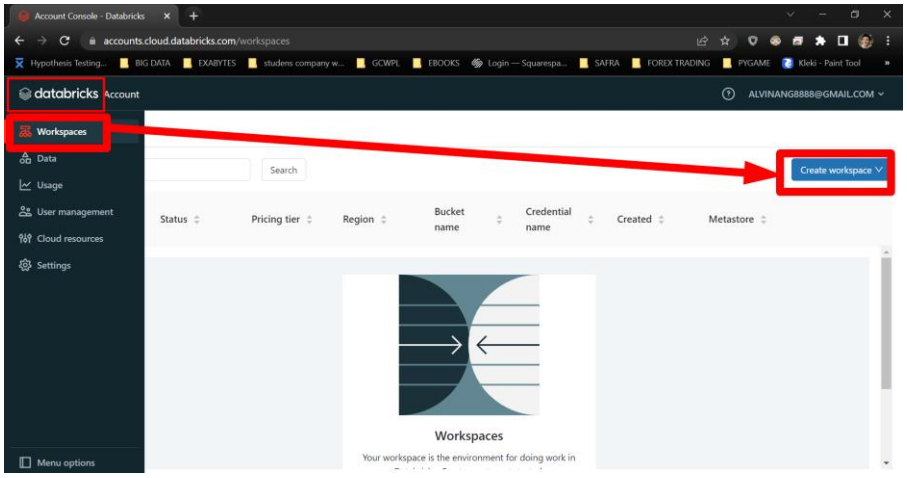
C. FINALIZE CONFIGURATIONS IN DATABRICKS

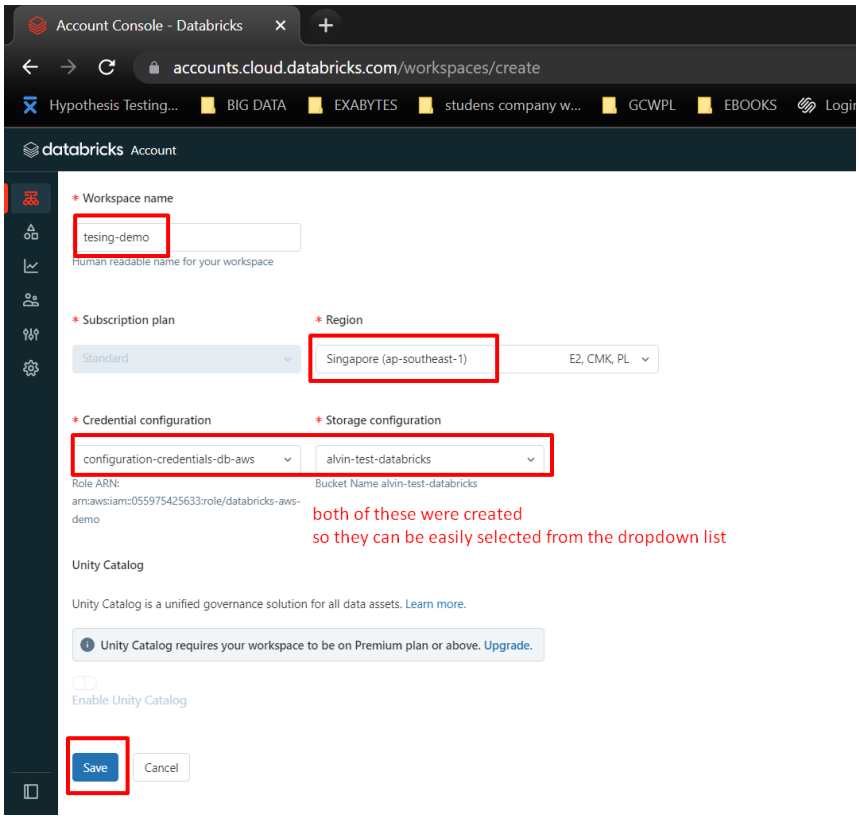


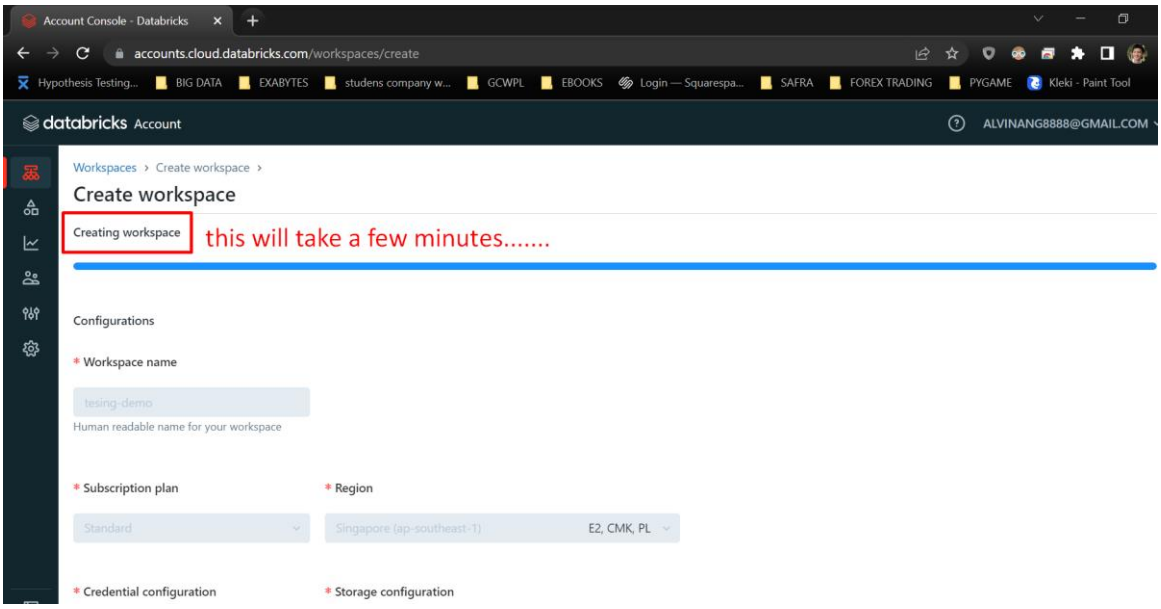


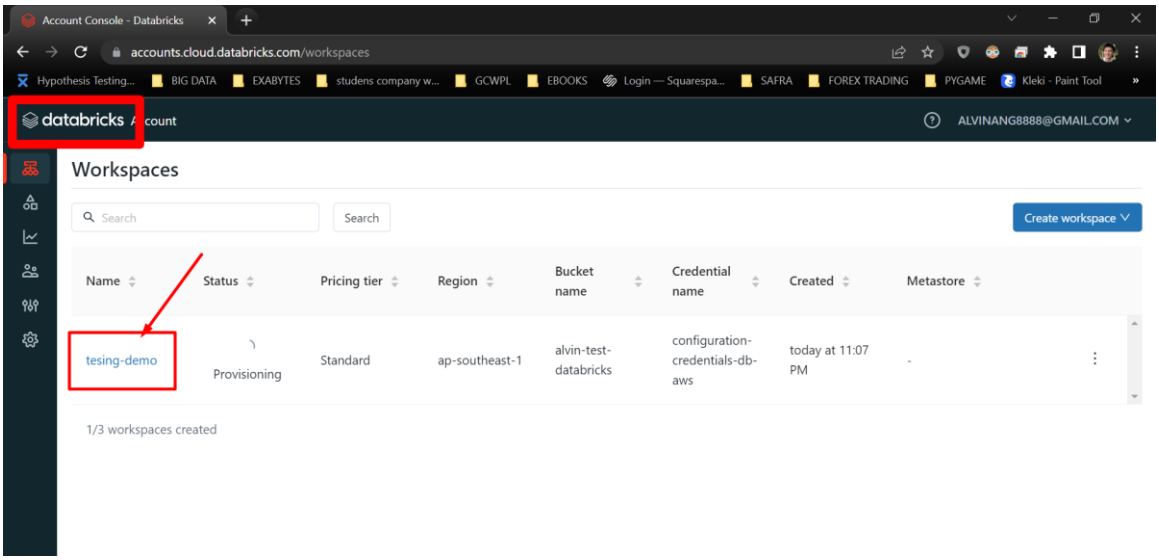
VI. CREATE WORKSPACE IN DATABRICKS

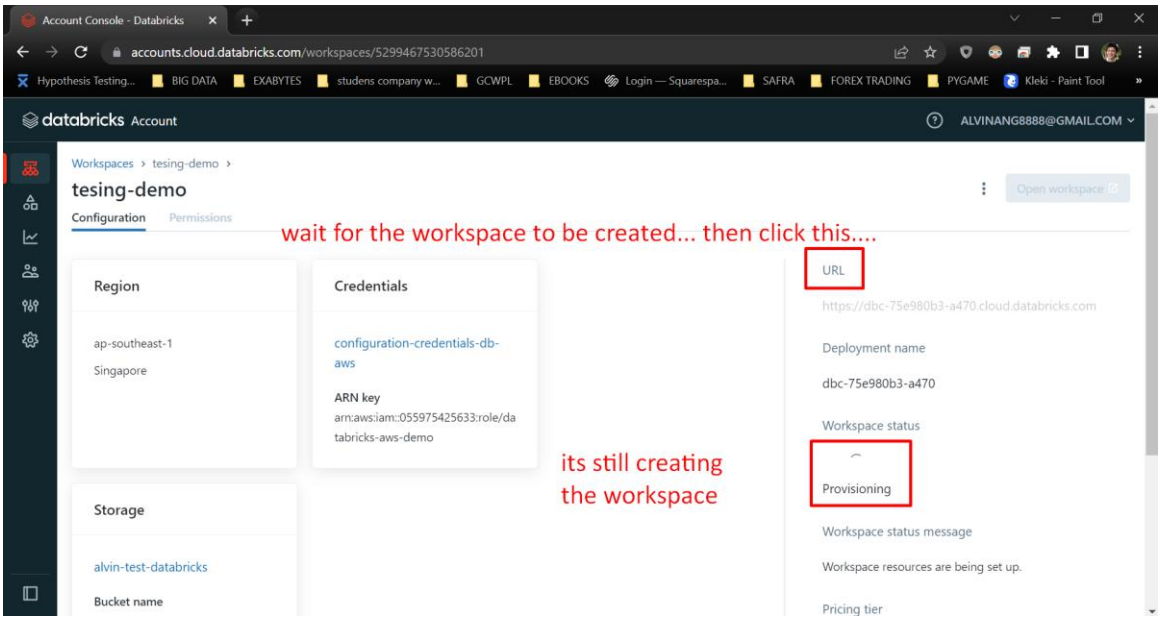
A. SETUP WORKSPACE

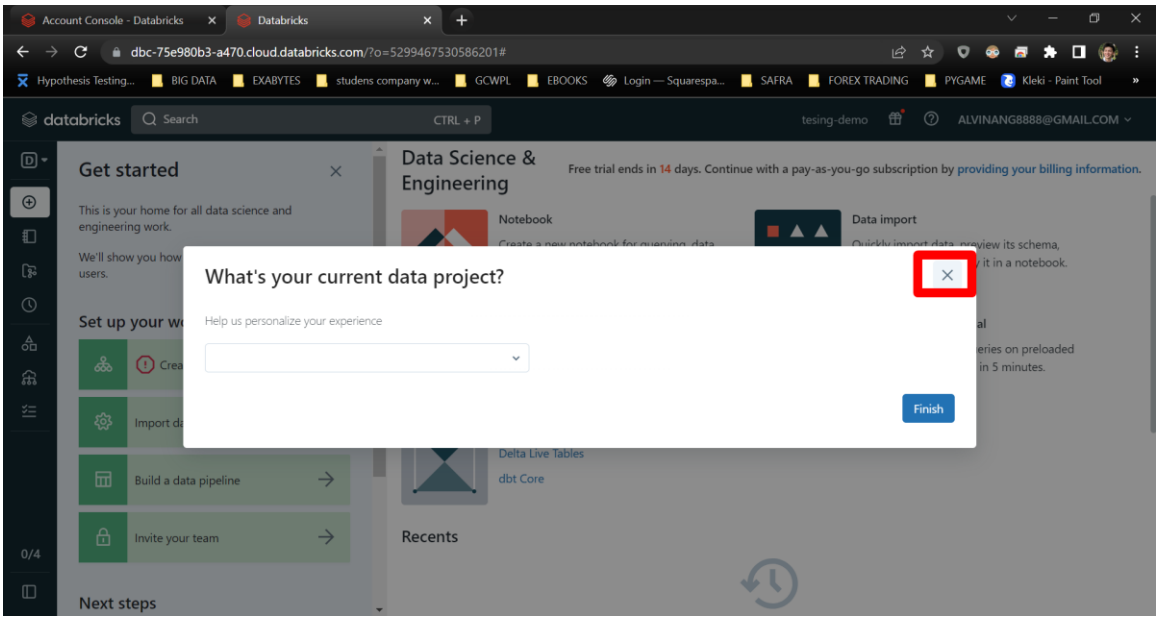




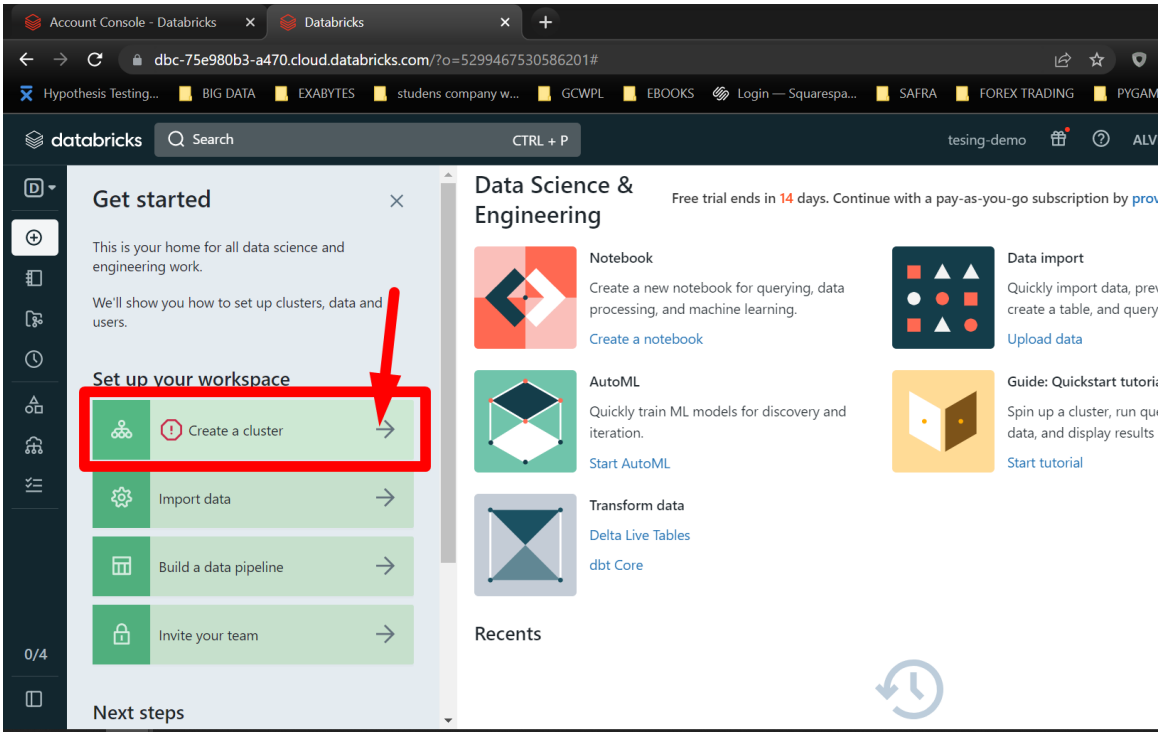


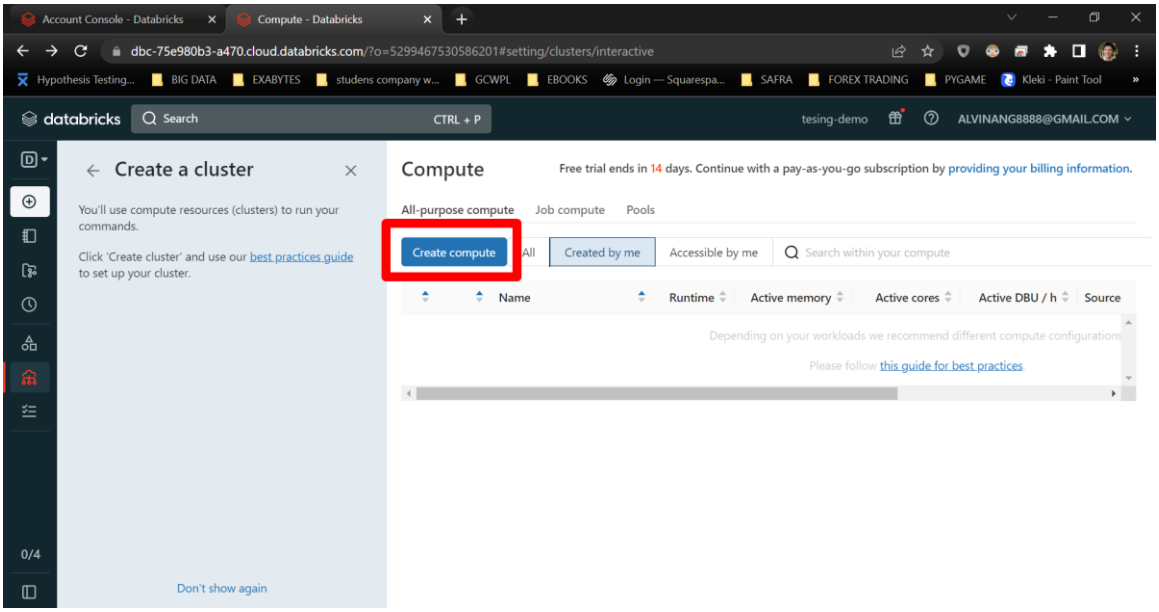




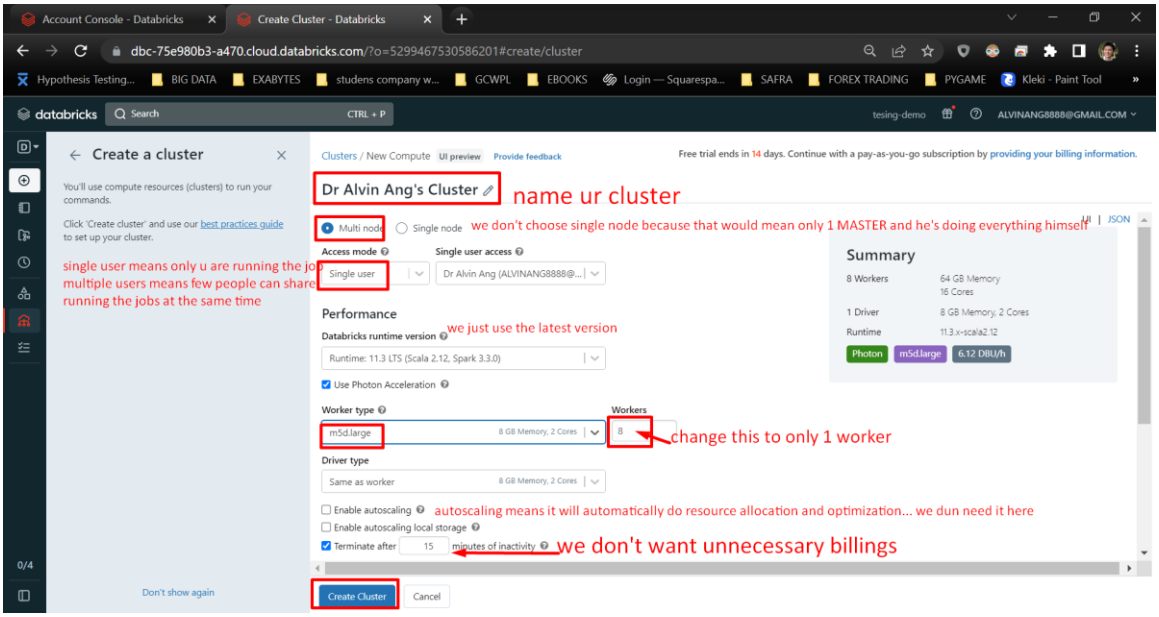


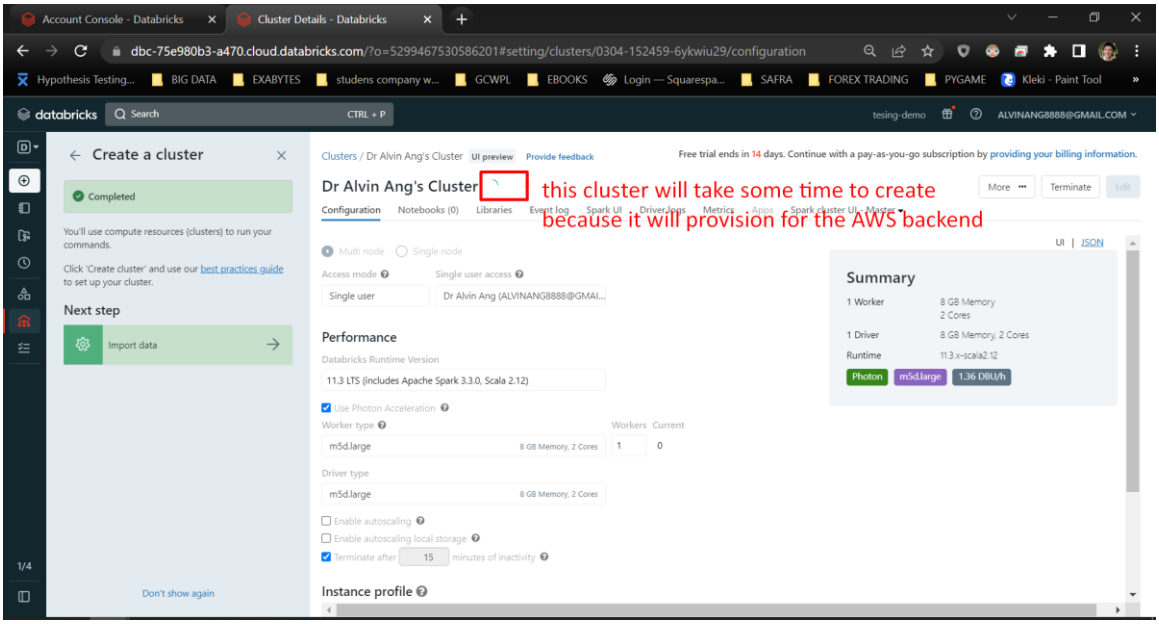
B. CREATE A CLUSTER



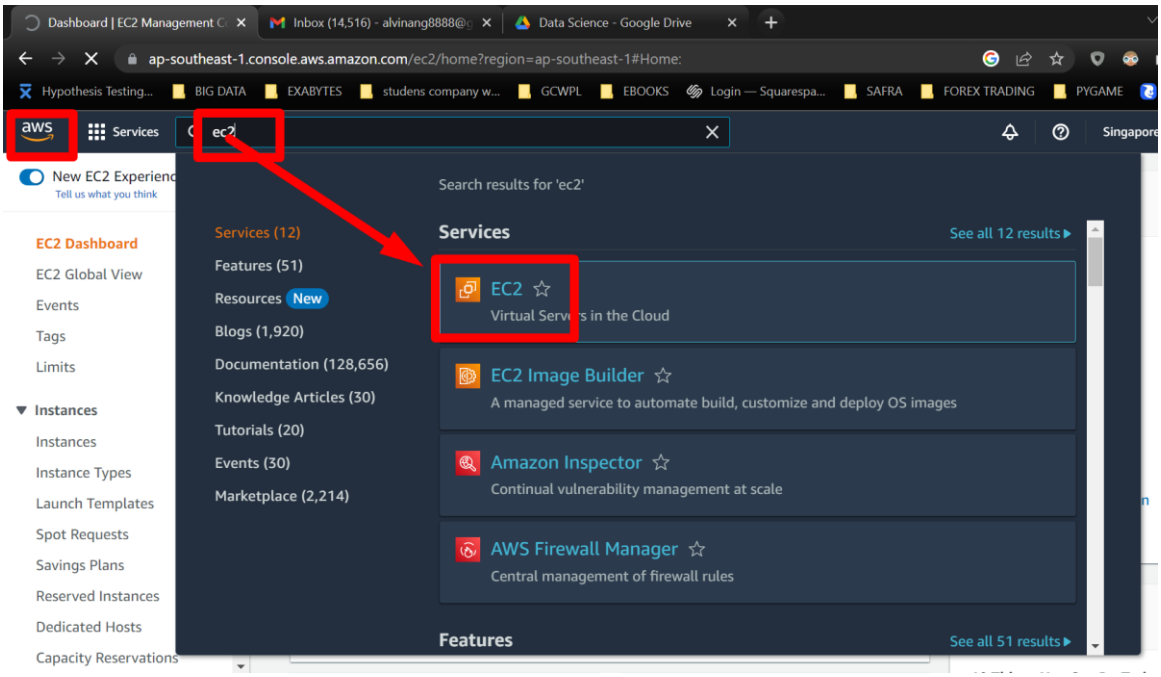


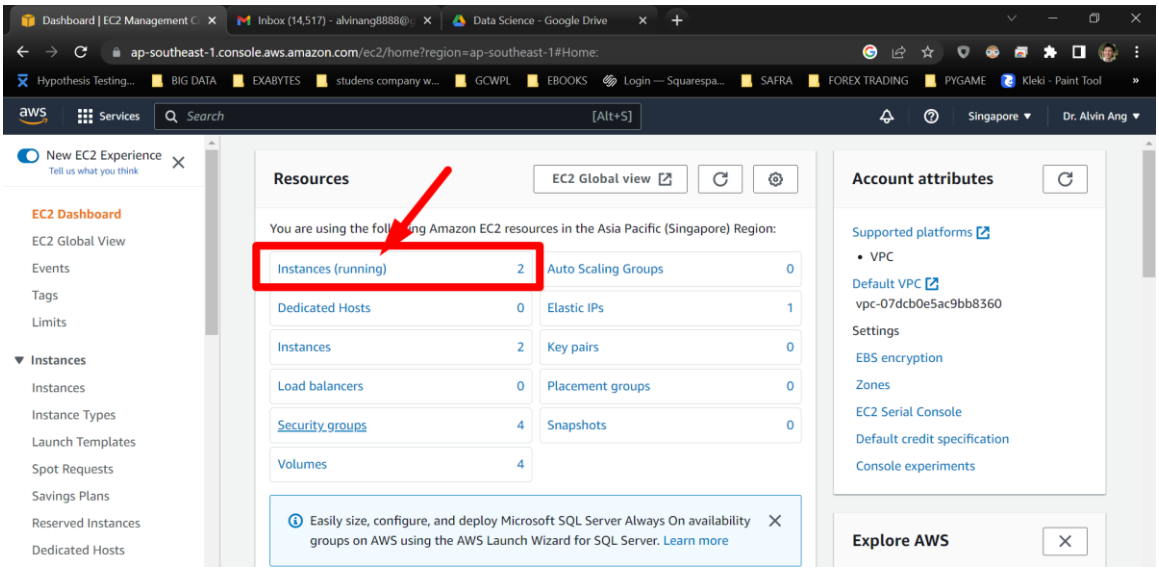
C. CONFIGURE CLUSTER

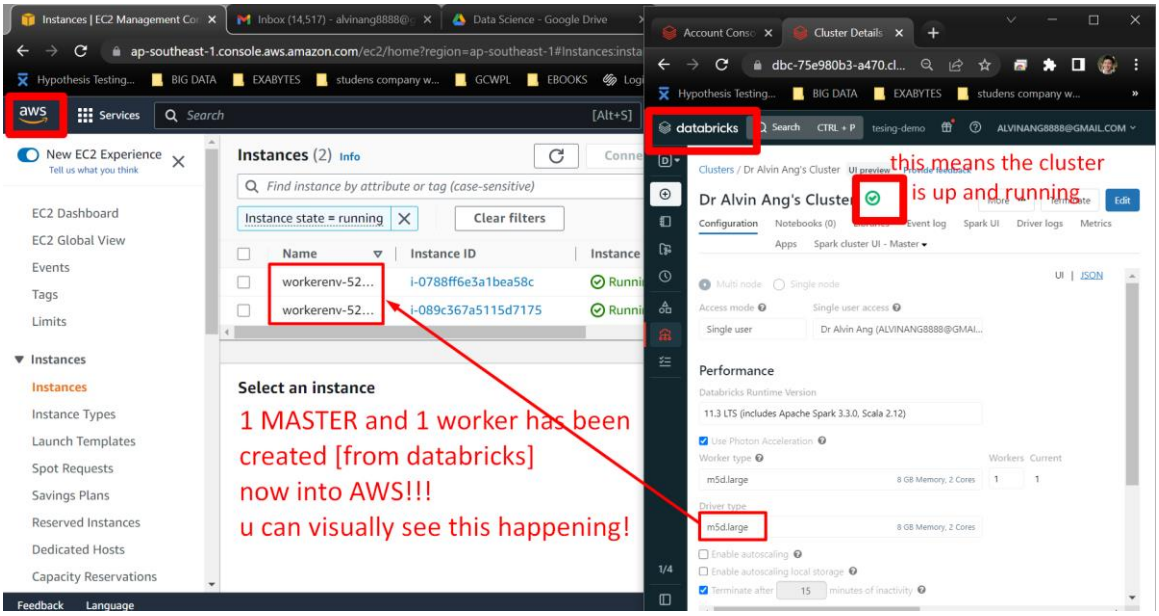


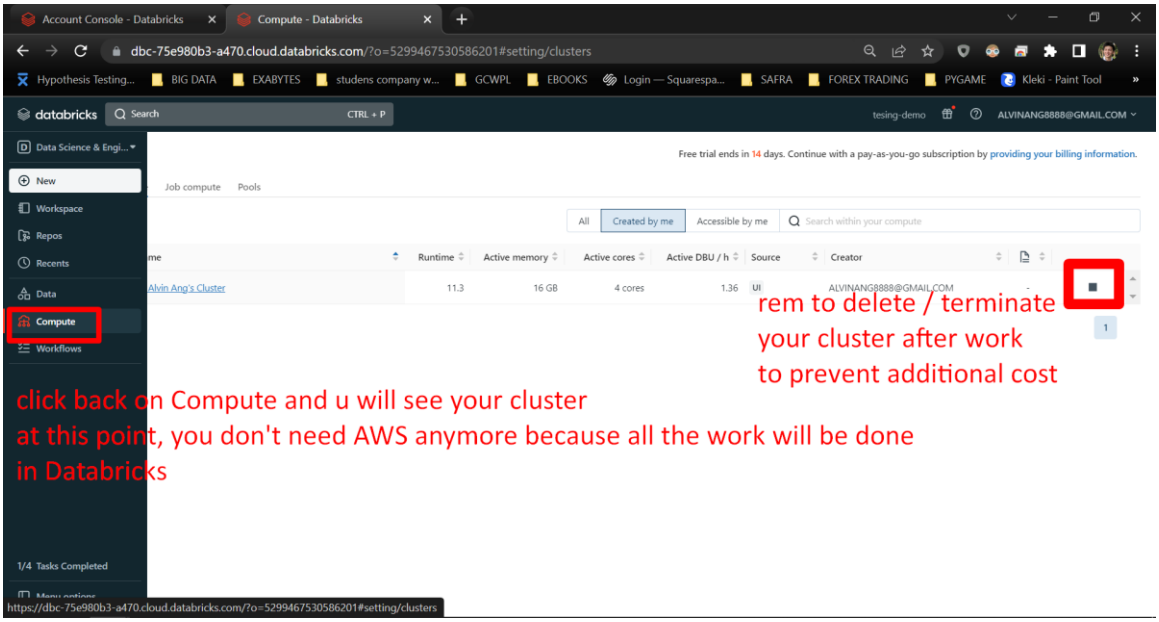


D. MEANWHILE, HEAD OVER TO AWS EC2



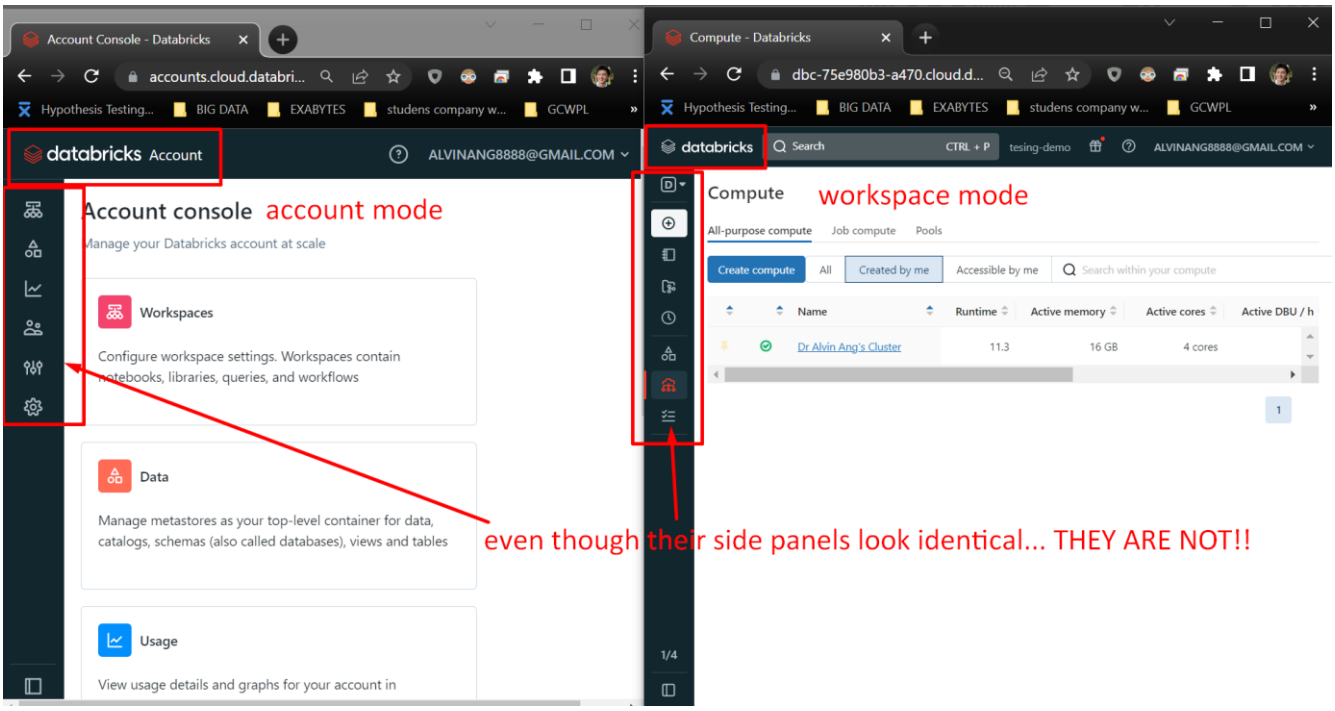




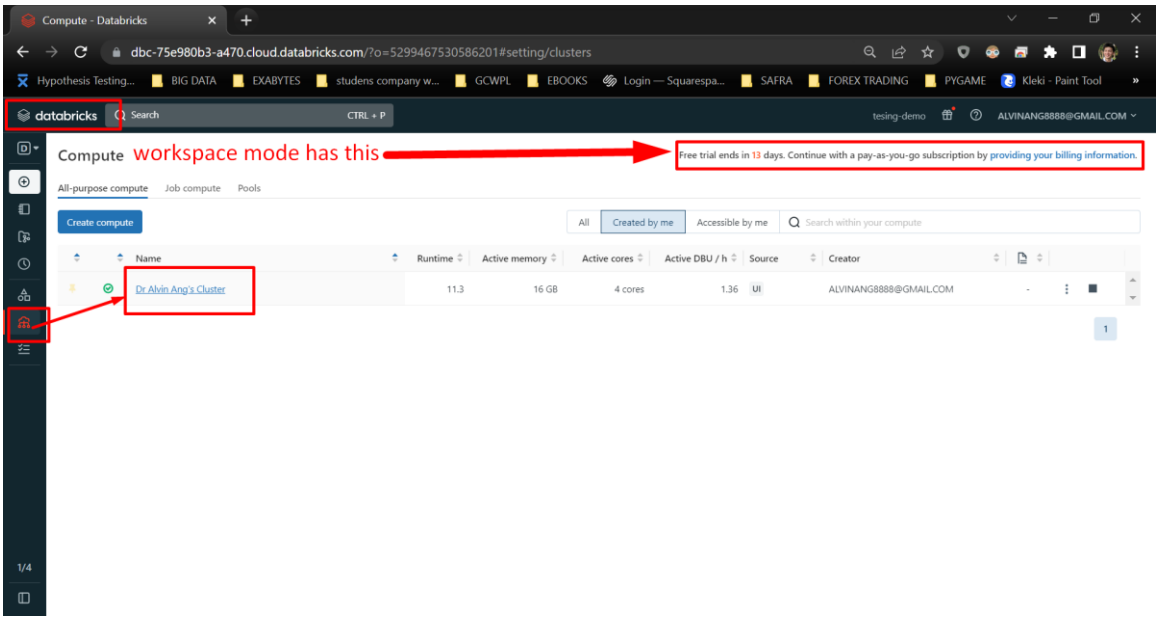


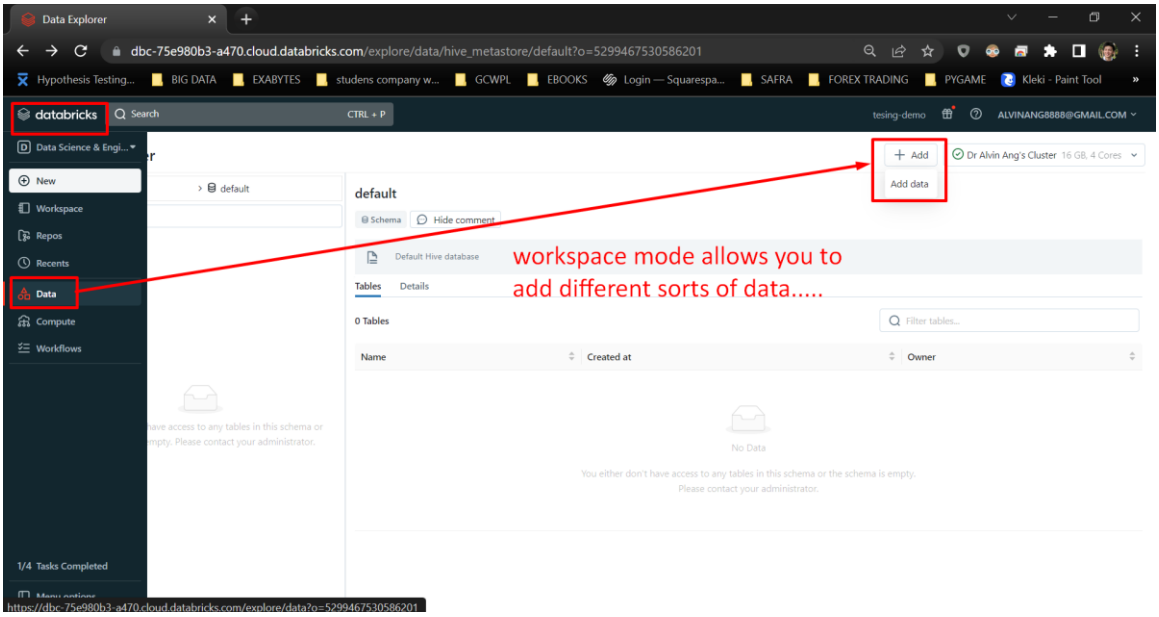
VII. IMPORTANT NOTE: DIFFERENCE BETWEEN ACCOUNT MODE VS WORKSPACE MODE

A. ACCOUNT MODE VS WORKSPACE MODE

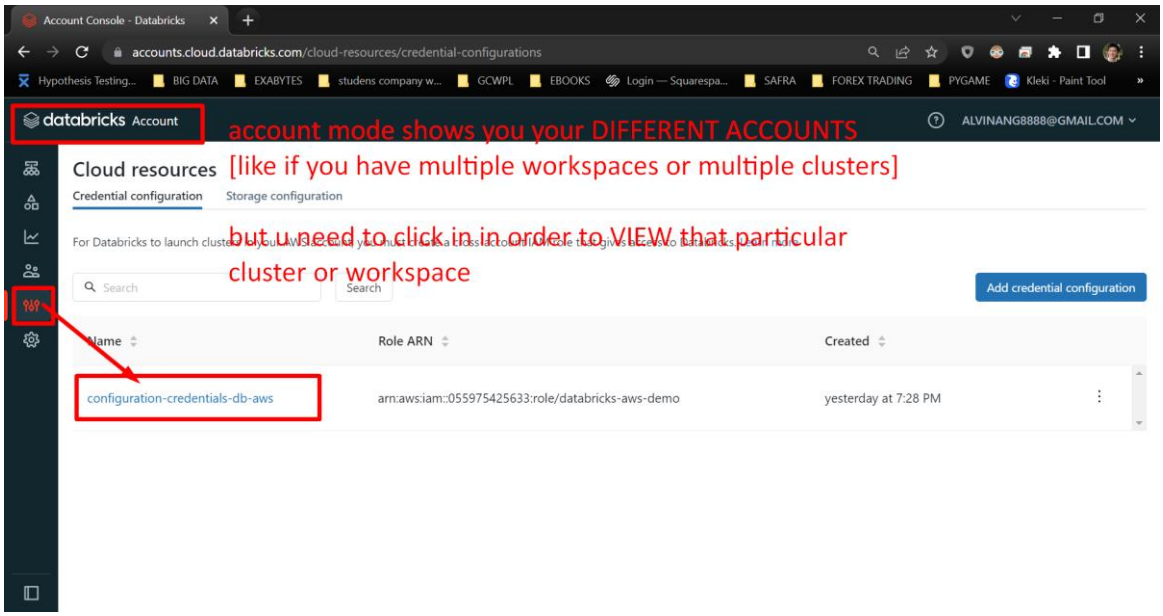


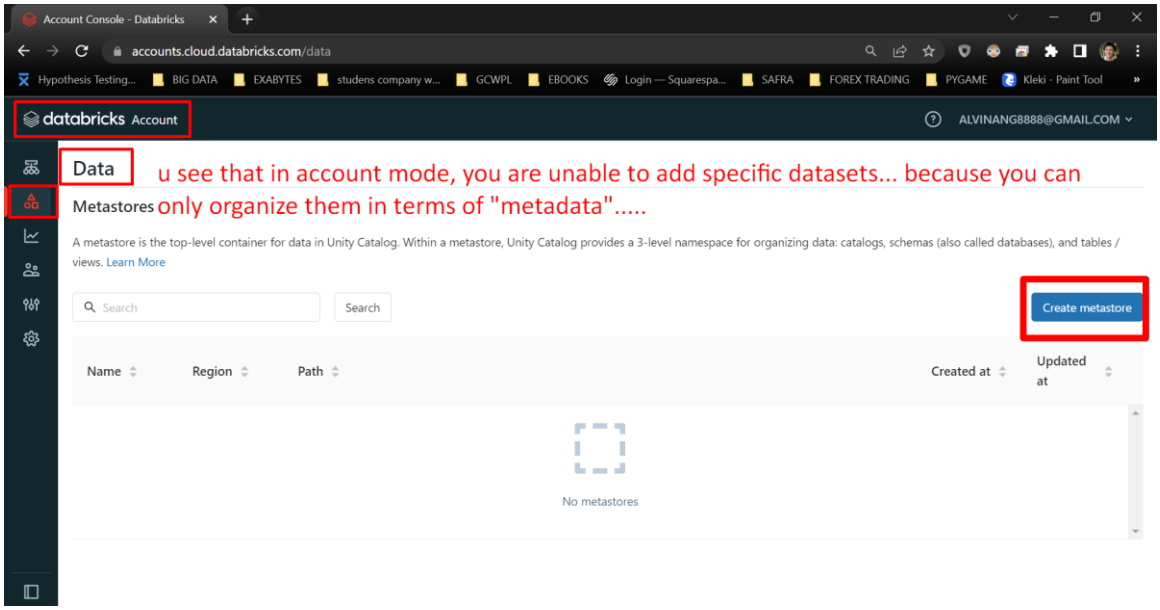
1. WORKSPACE MODE



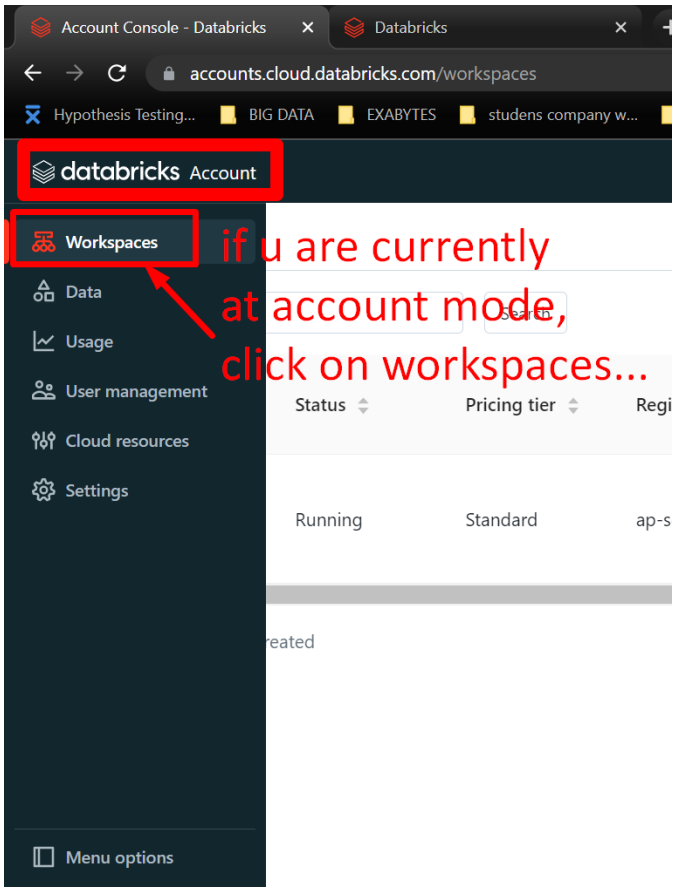


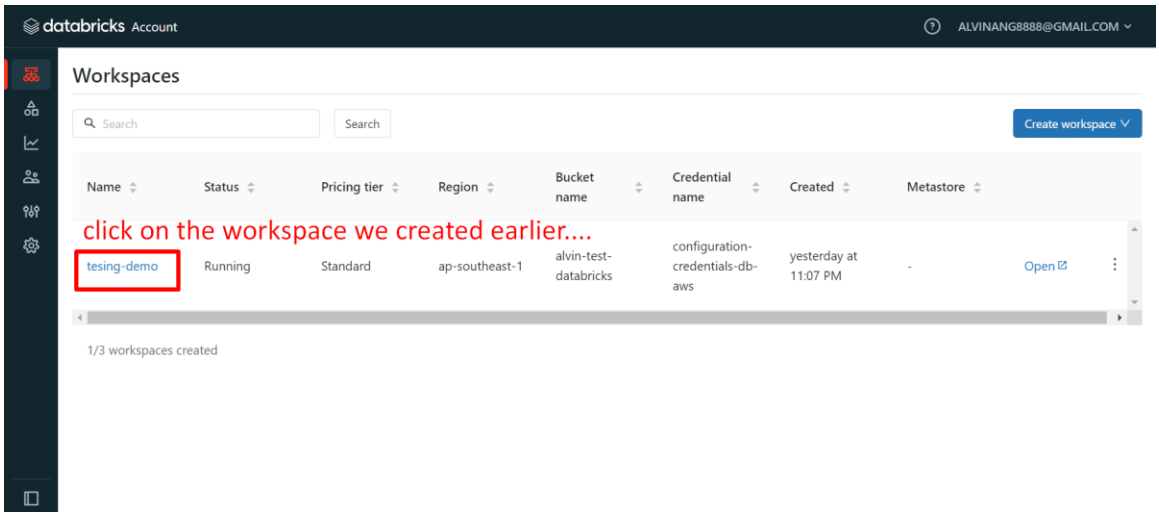
2. ACCOUNT MODE

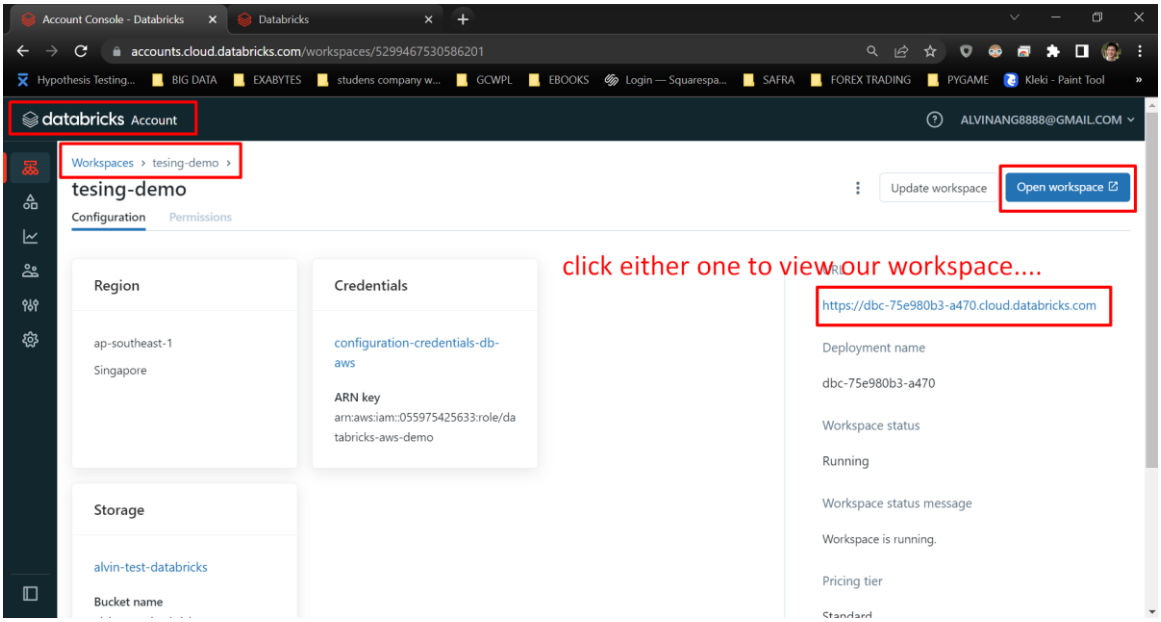


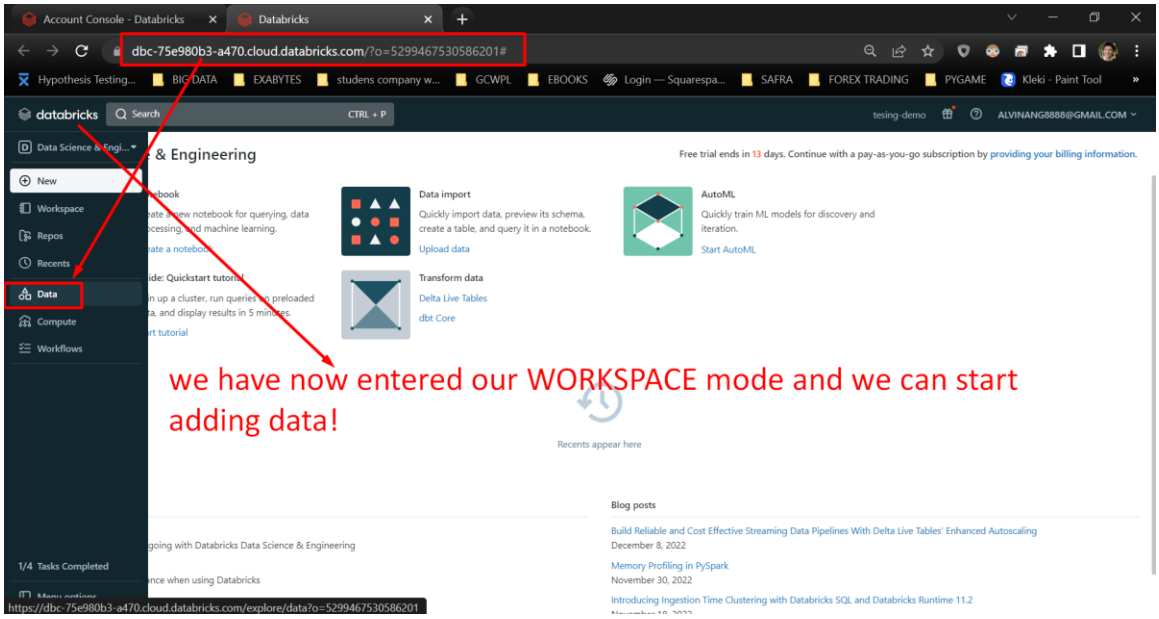


B. YOU MUST MOVE FROM ACCOUNT MODE TO WORKSPACE MODE

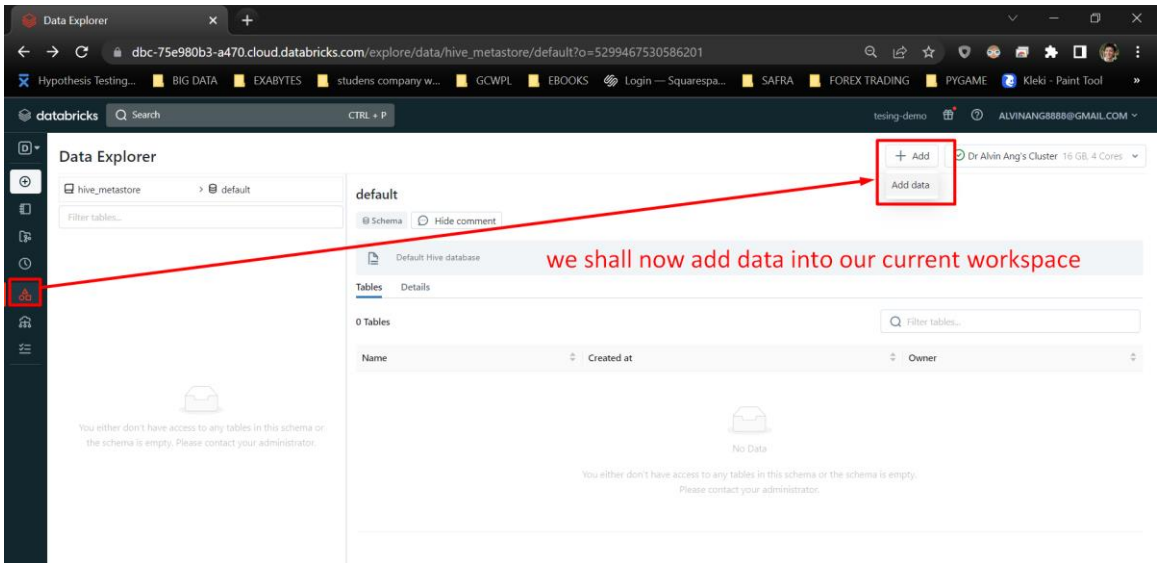


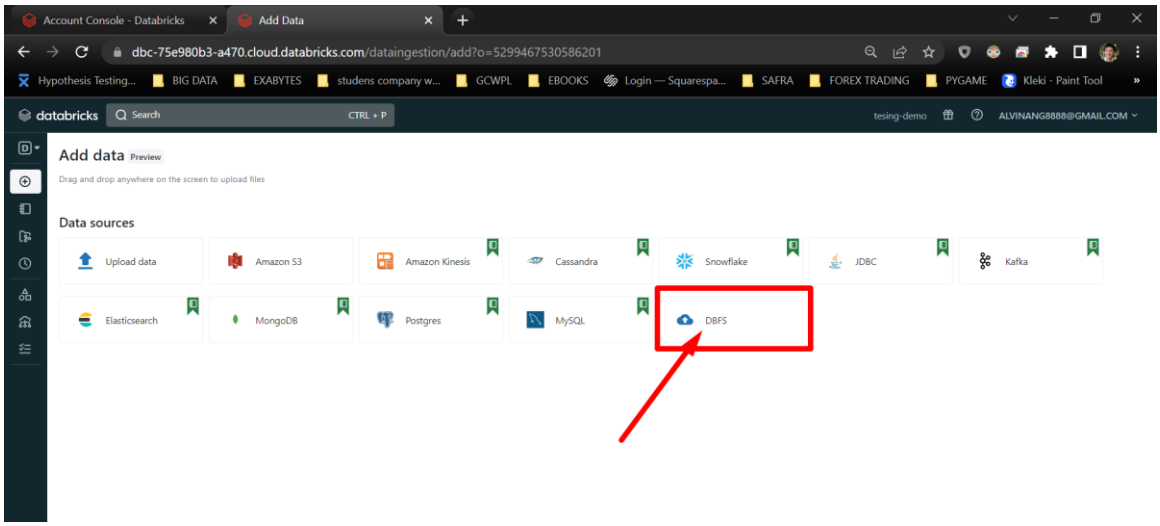


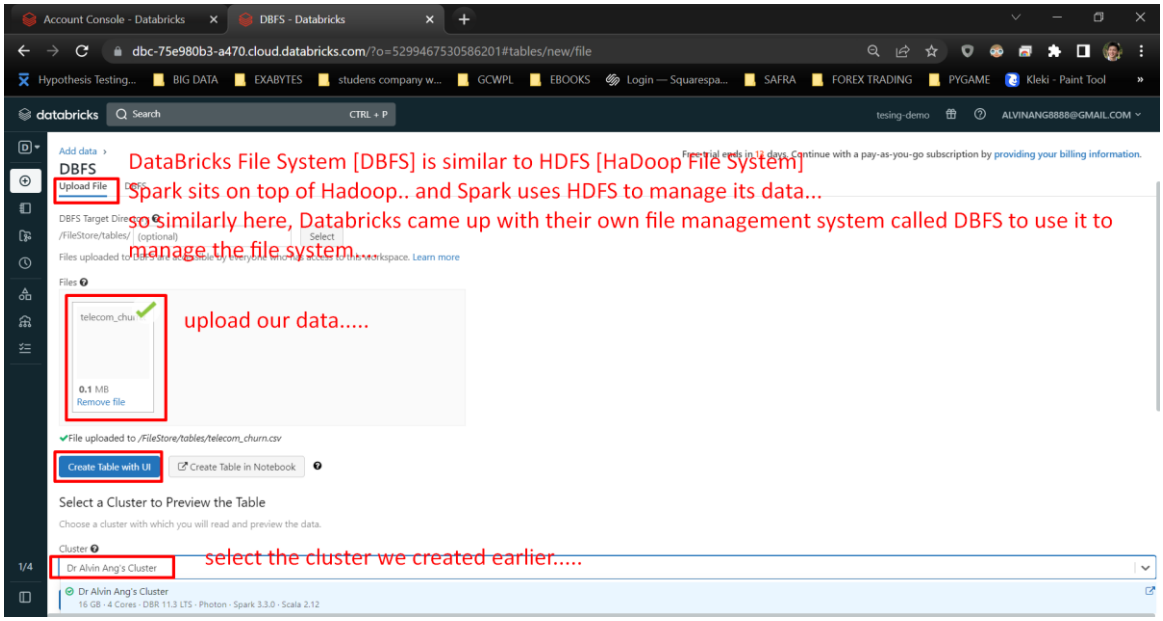


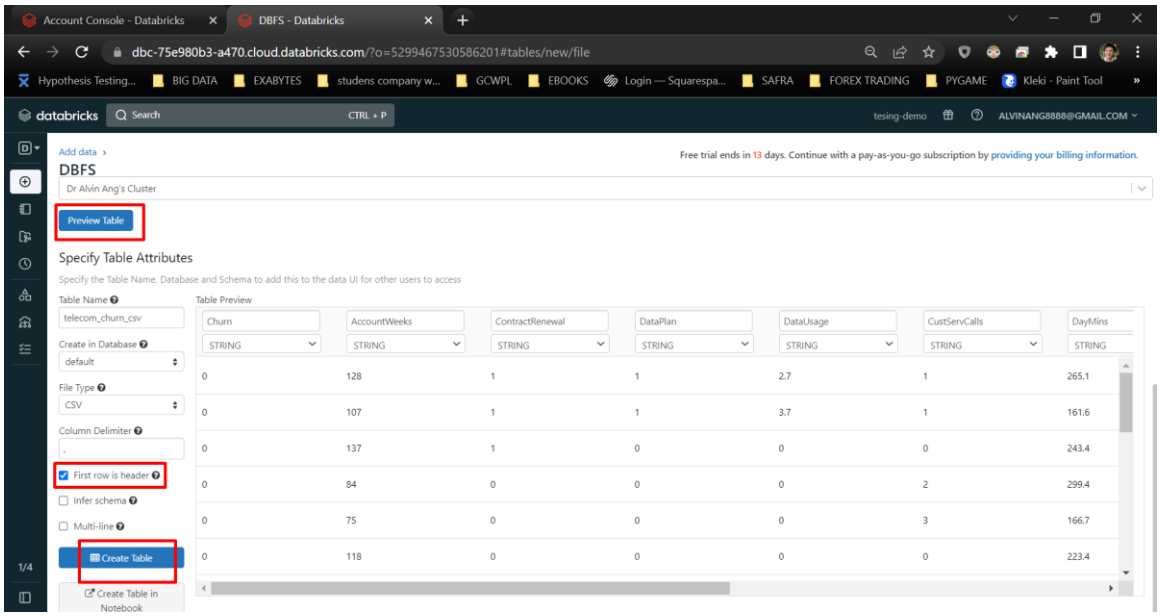


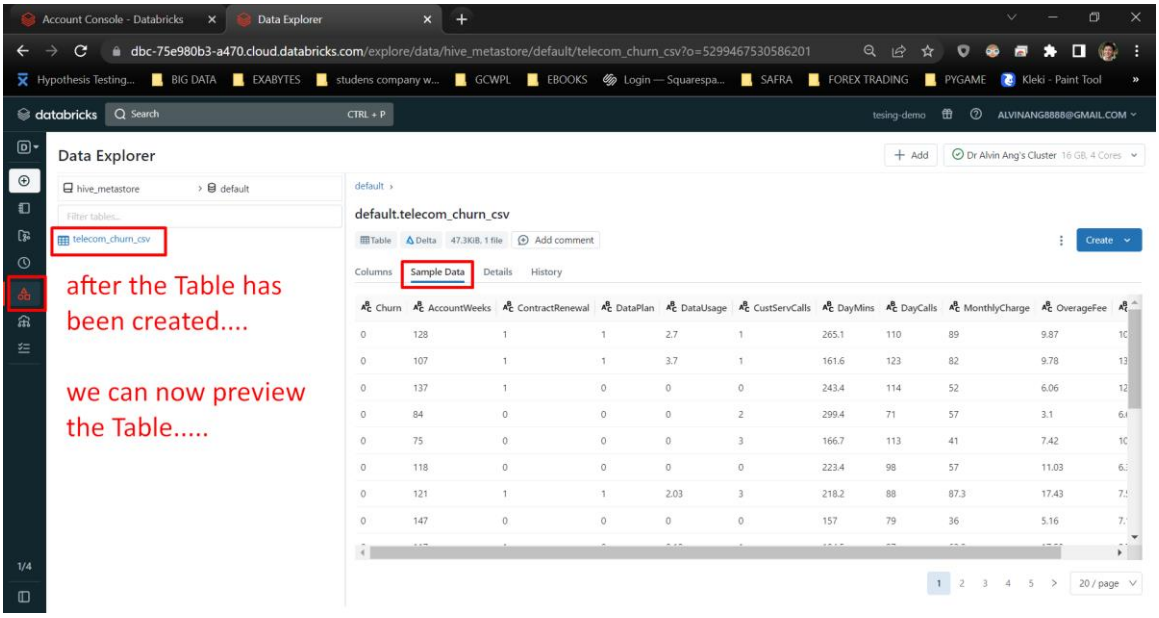
VIII. CREATE TABLES IN DATABRICKS

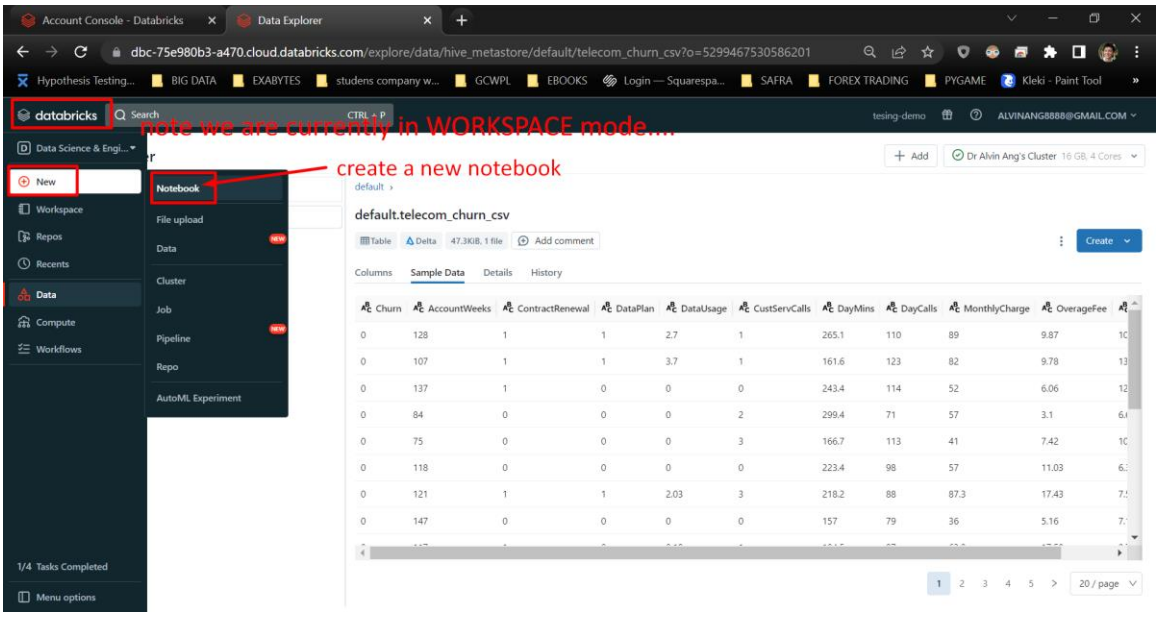


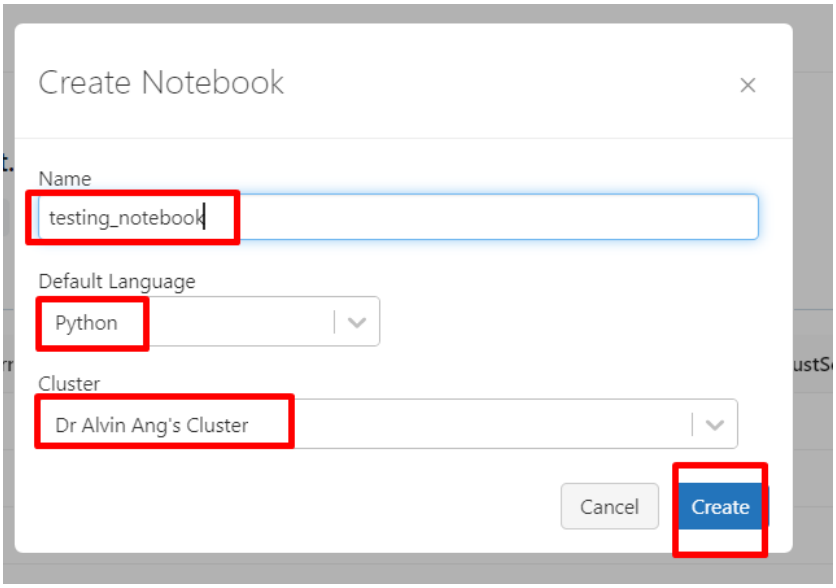




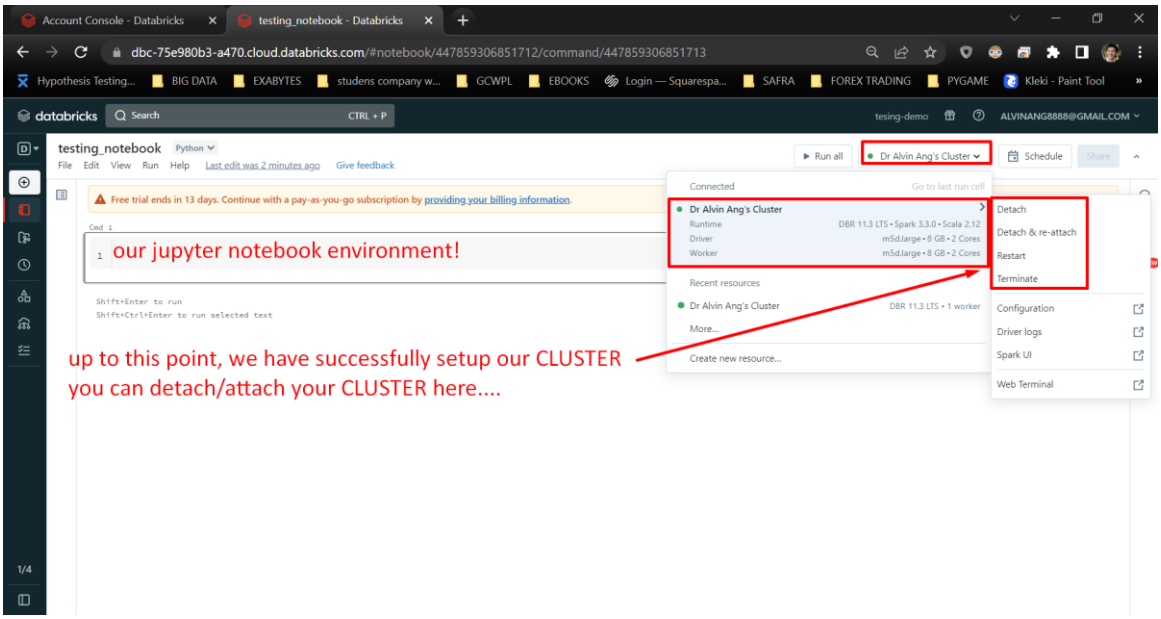




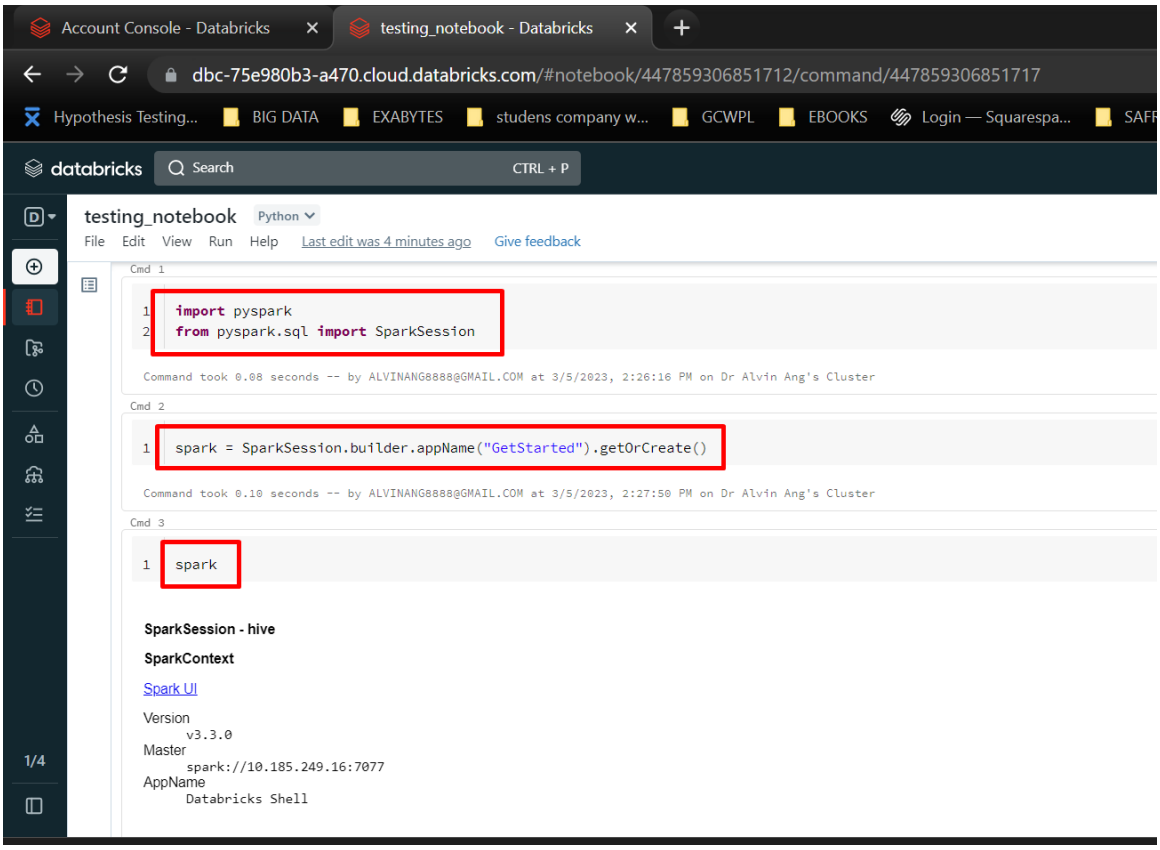




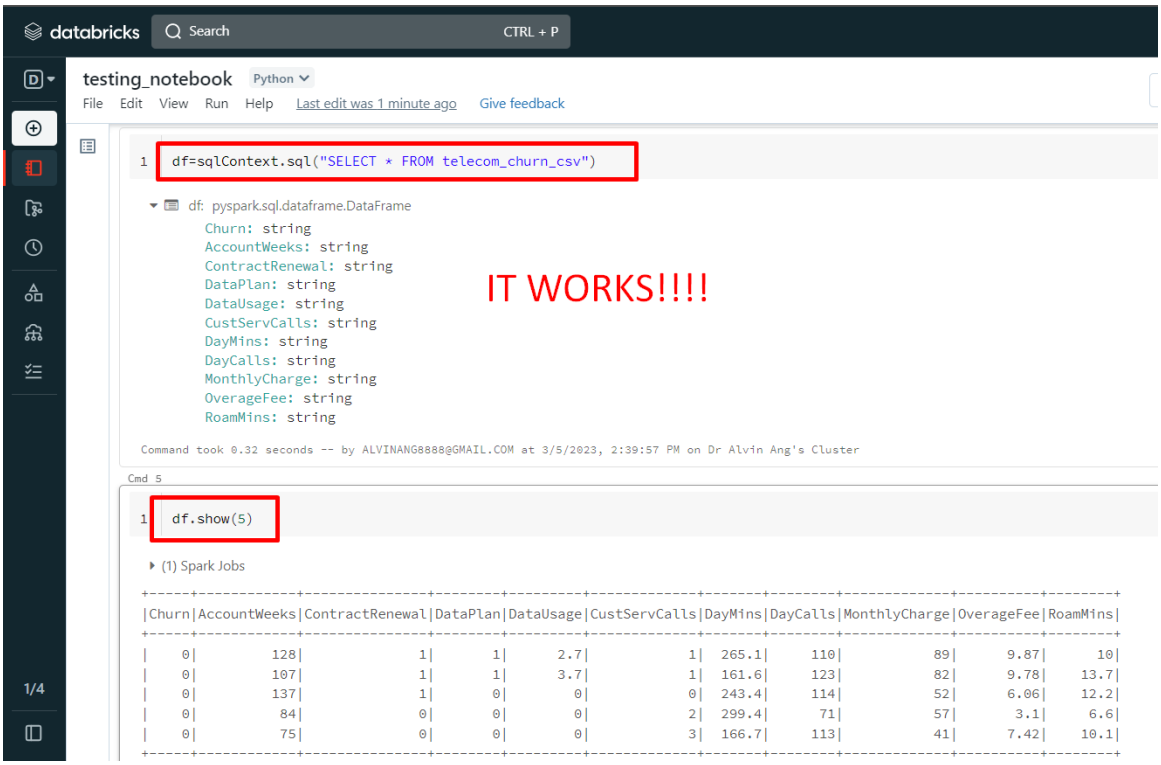
A. ATTACH / DETACH CLUSTER...



B. START SPARK SESSION

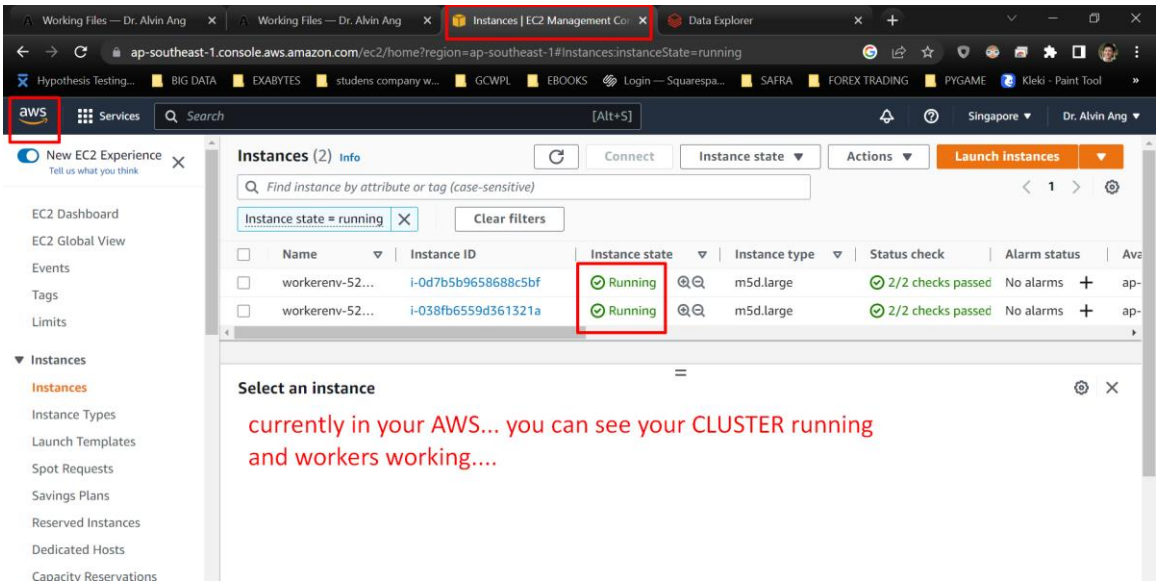


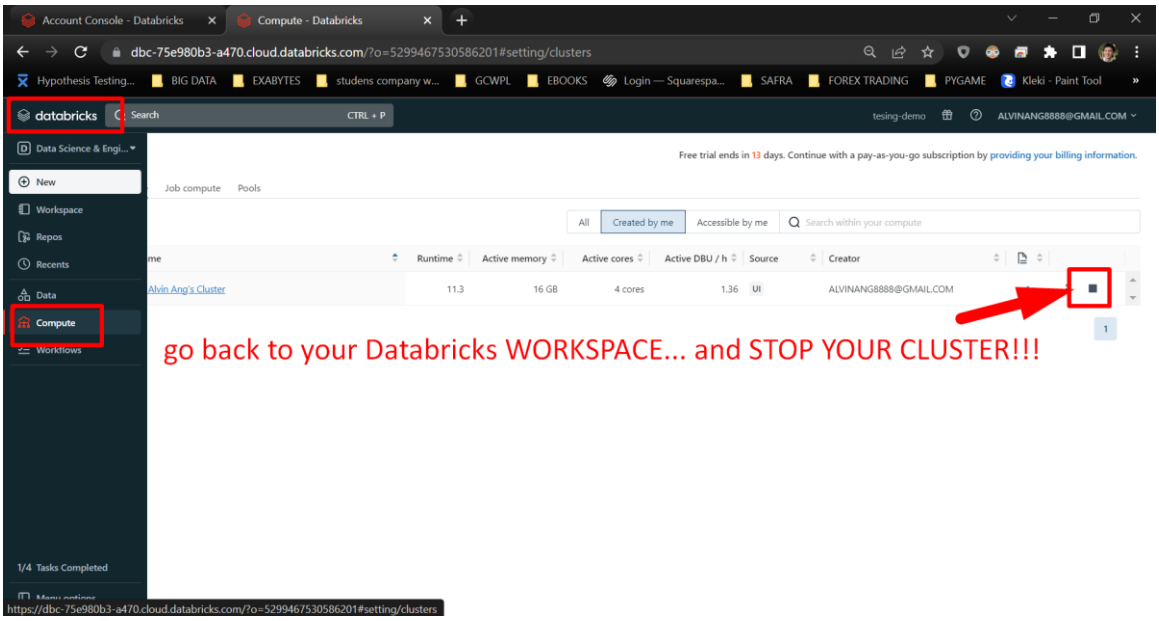
C. TEST SOME CODE….

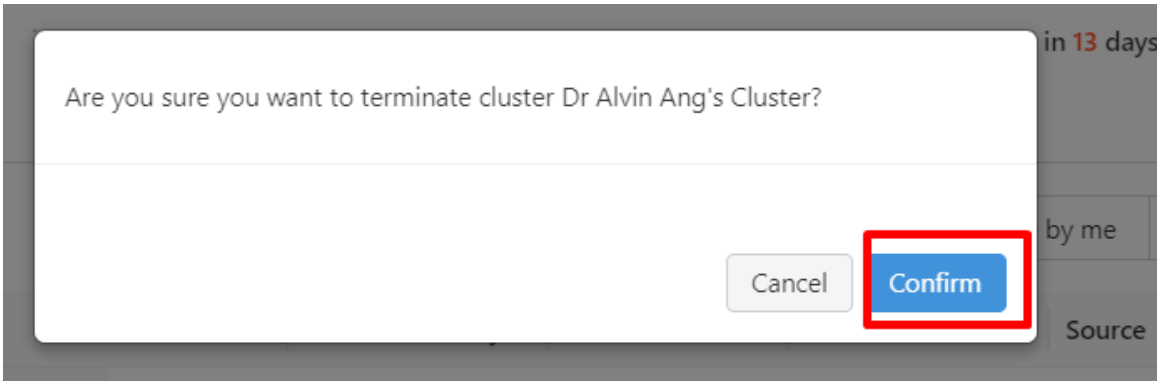


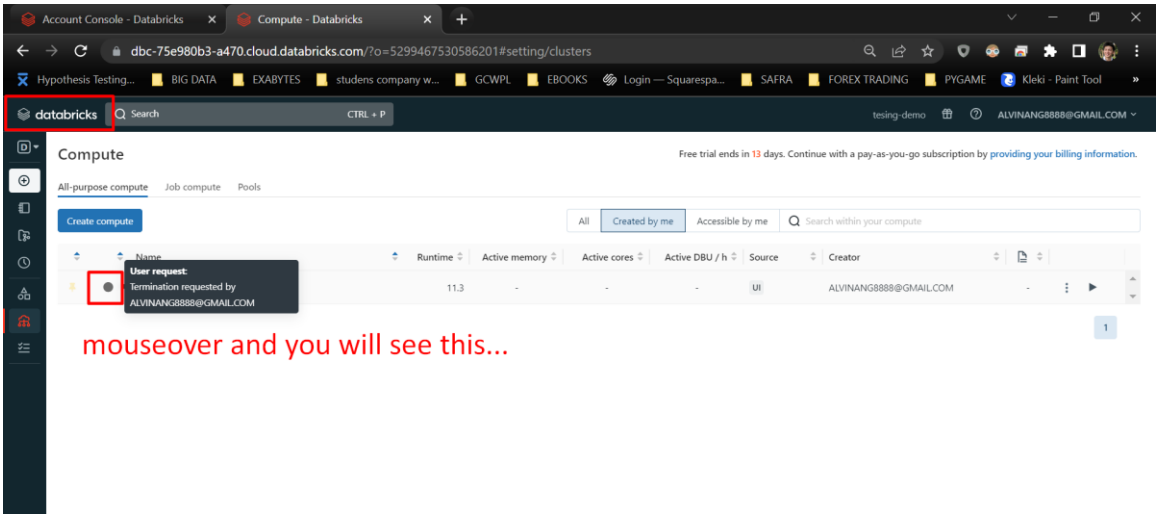
X. VERY IMPORTANT: SHUT DOWN YOUR CLUSTER!!!

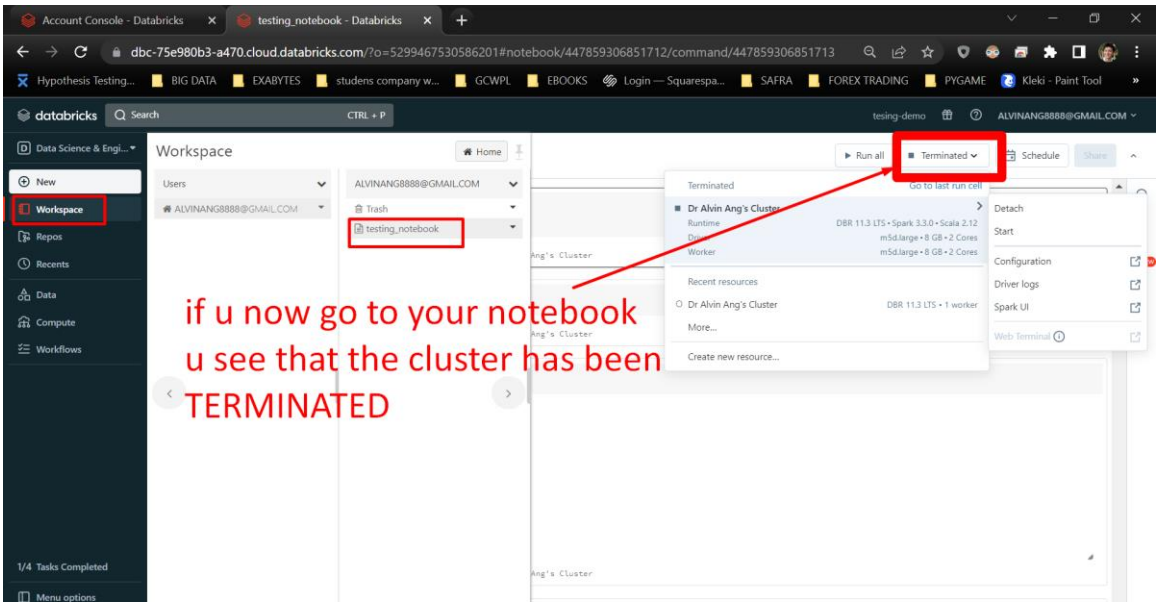
To prevent further charges…

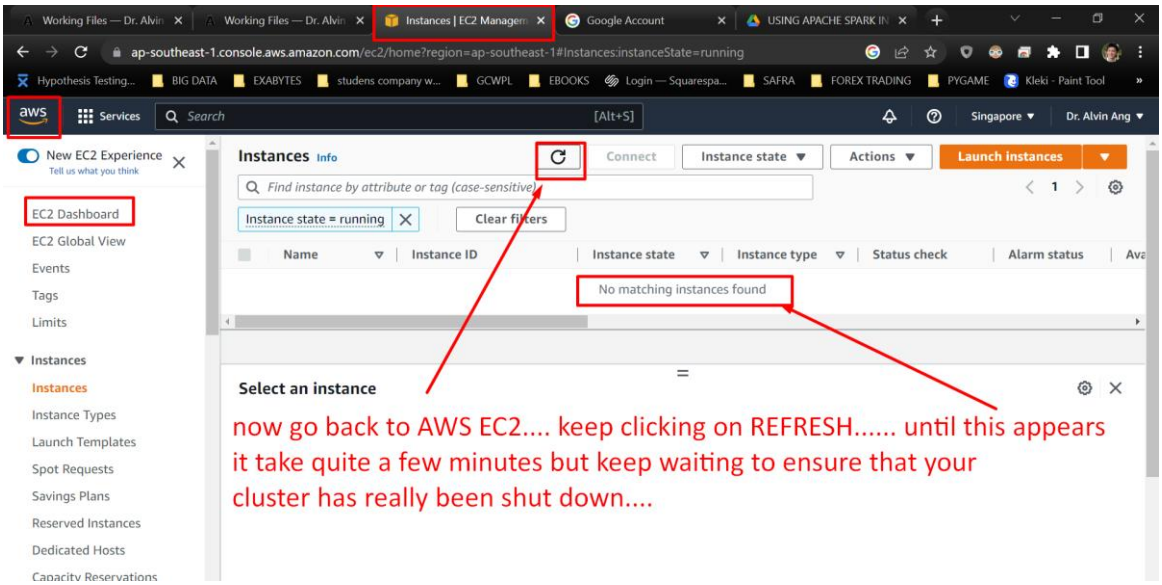




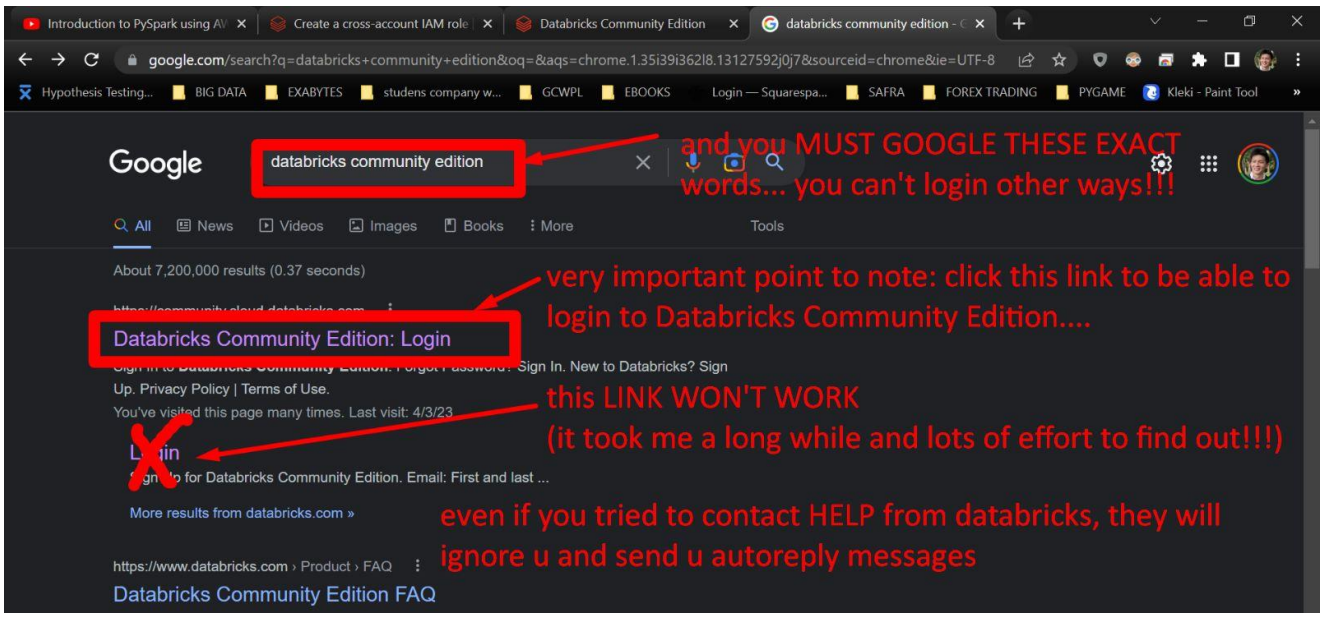




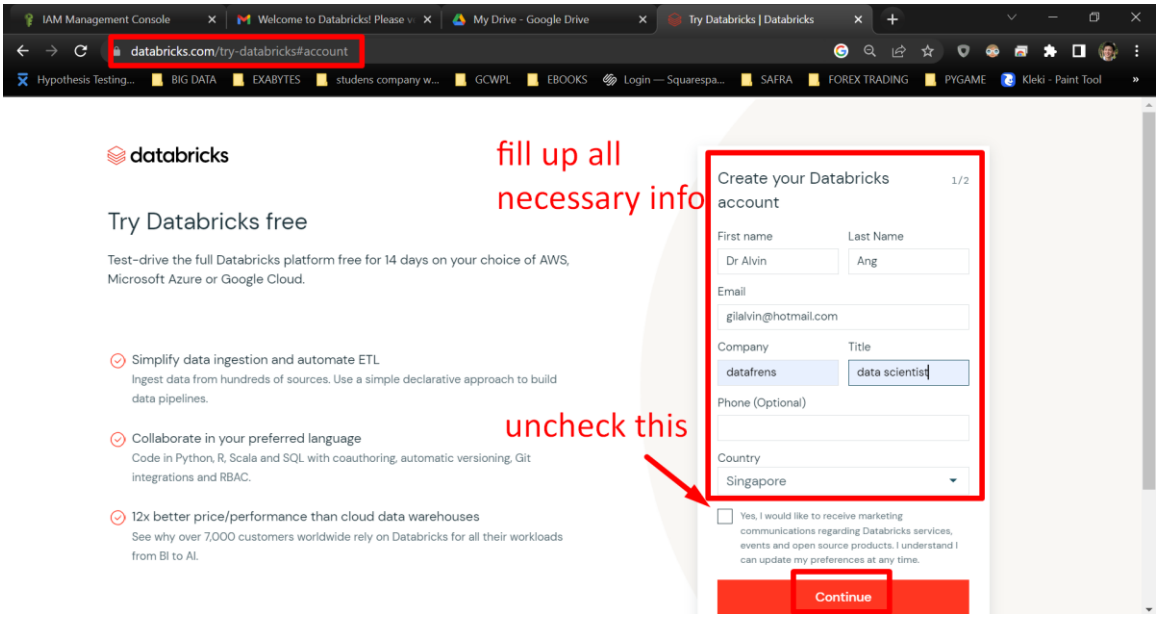


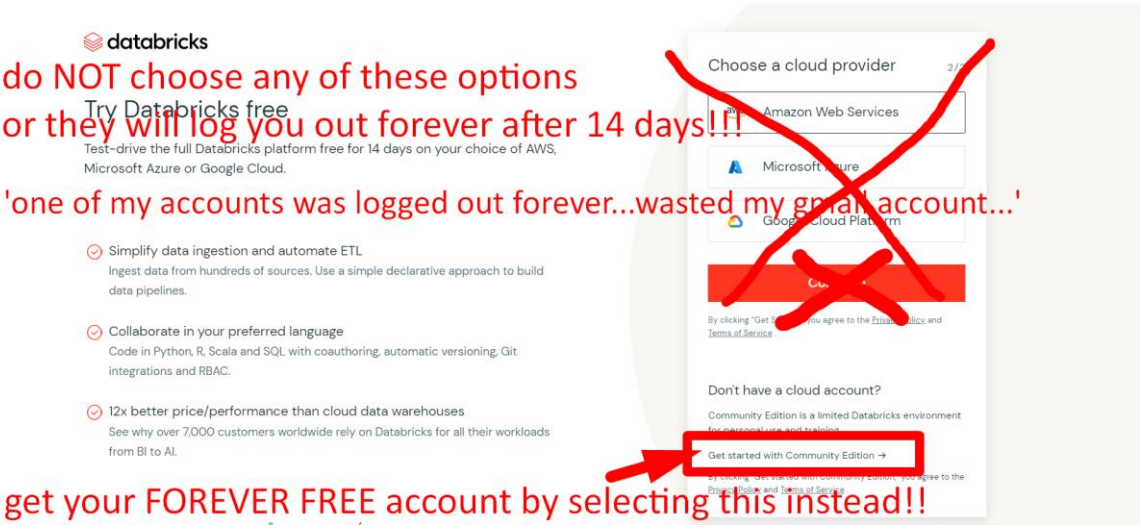


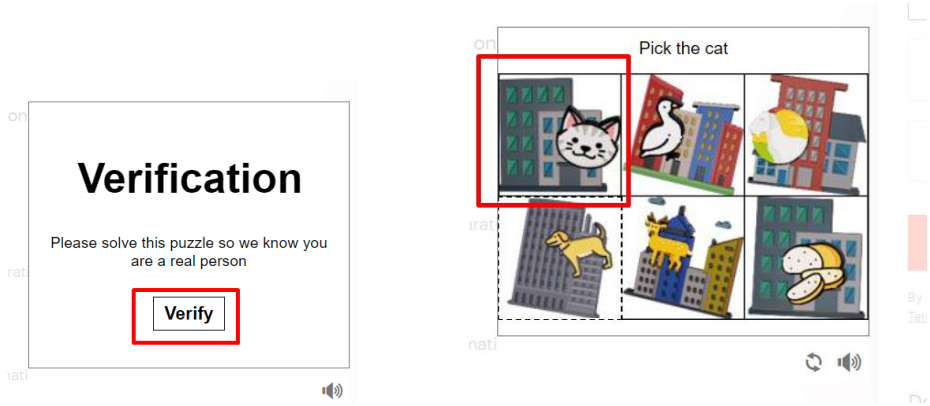
XI. APPENDIX I: SIGNING UP WITH DATABRICK COMMUNITY EDITION [FREE FOREVER BUT WE WON'T BE USING THIS OPTION IN THIS MANUSCRIPT]

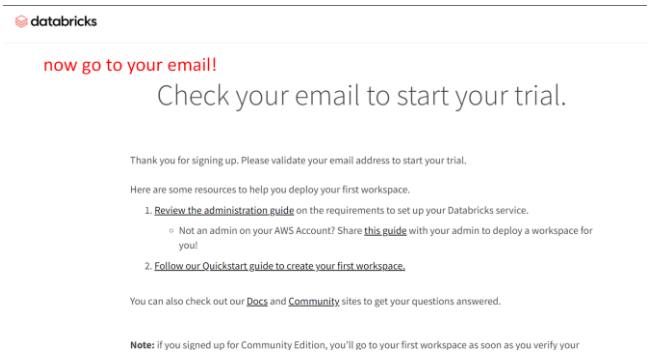


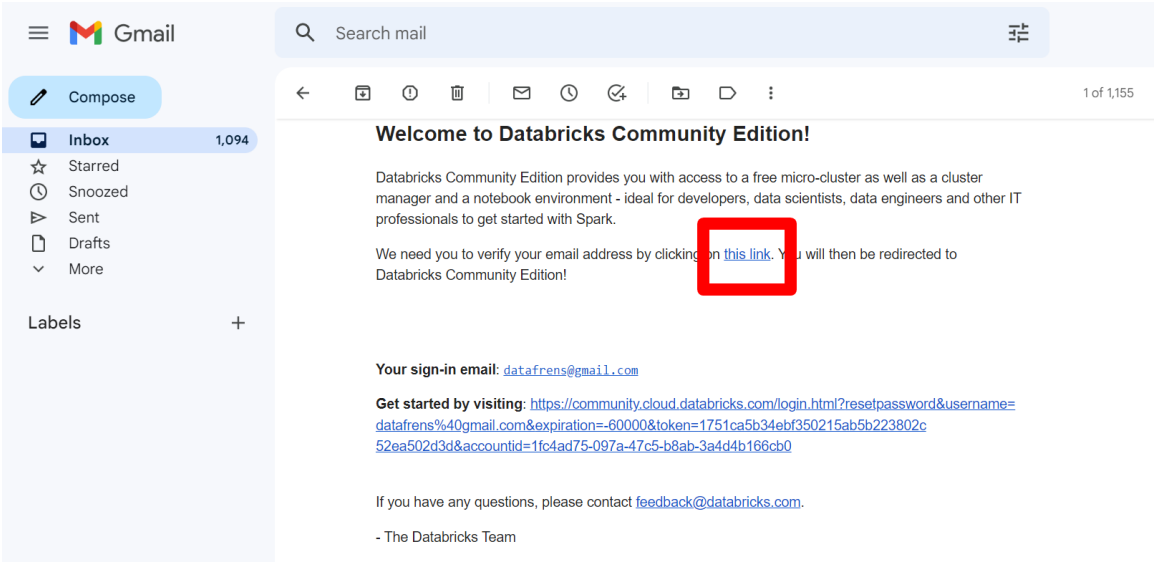
https://www.databricks.com/try-databricks#account

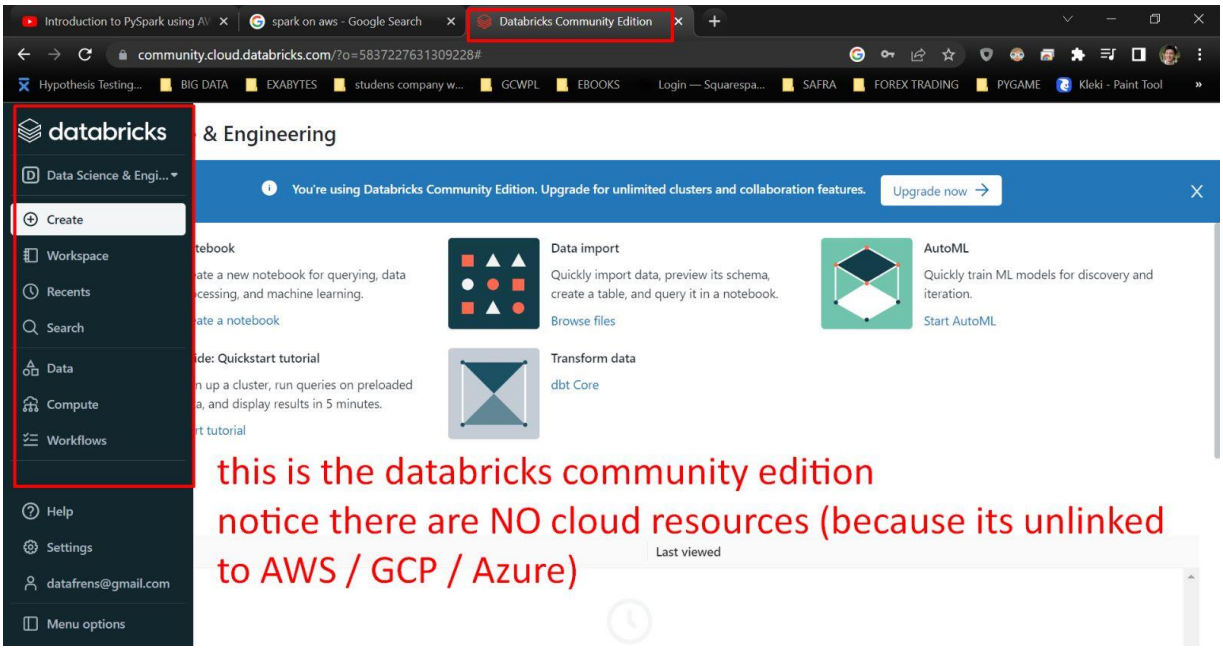












Dr. Alvin Ang earned his Ph.D., Masters and Bachelor degrees from NTU, Singapore. He is a scientist, entrepreneur, as well as a personal/business advisor. More about him at www.AlvinAng.sg.