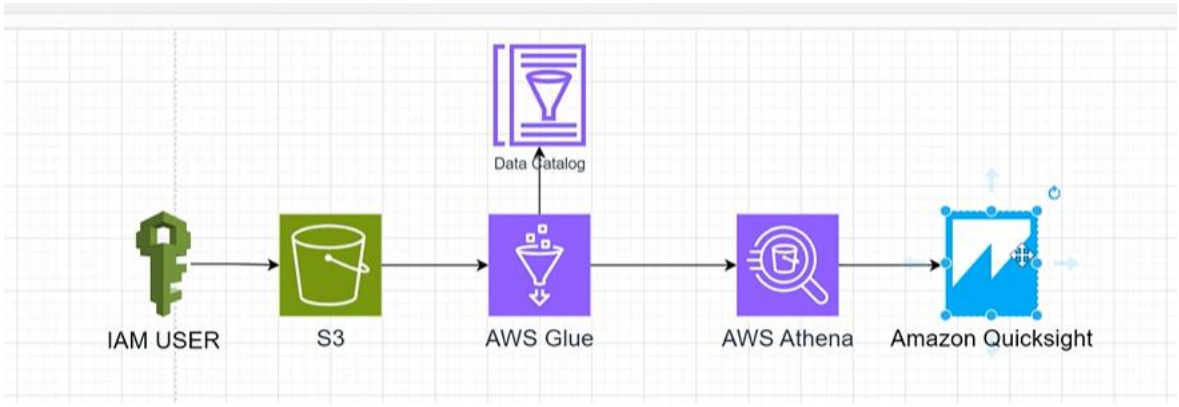
End to End AWS Data Engineering Project

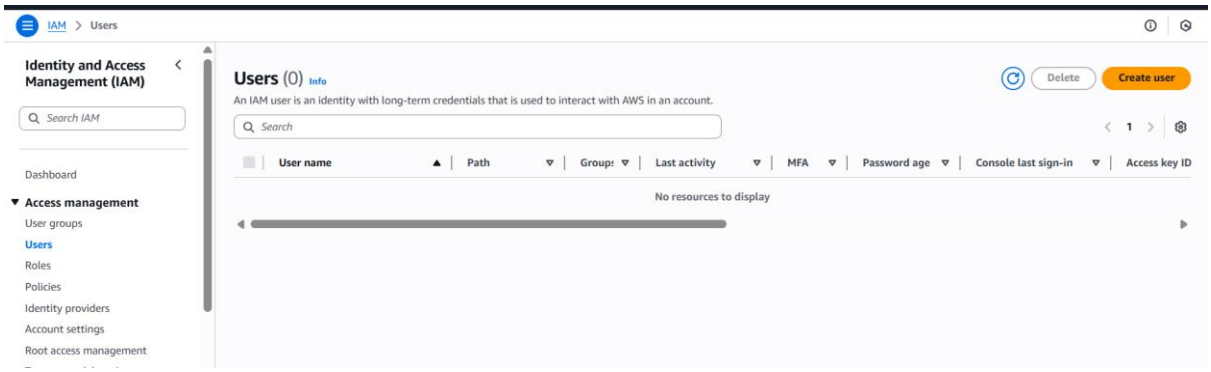
Data Pipeline design:

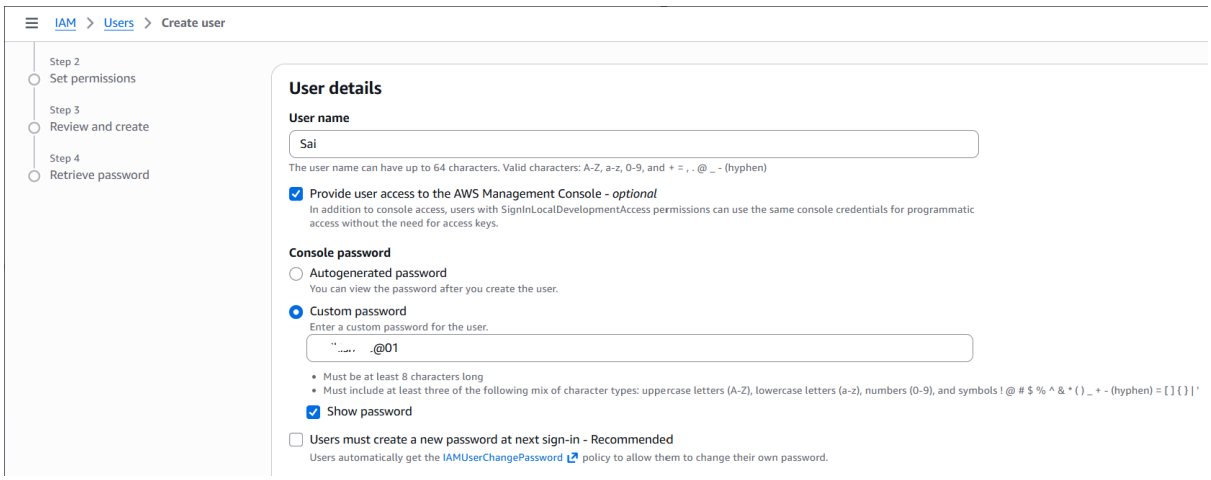


IAM (identity and access management) where you can create users, groups, policies etc.

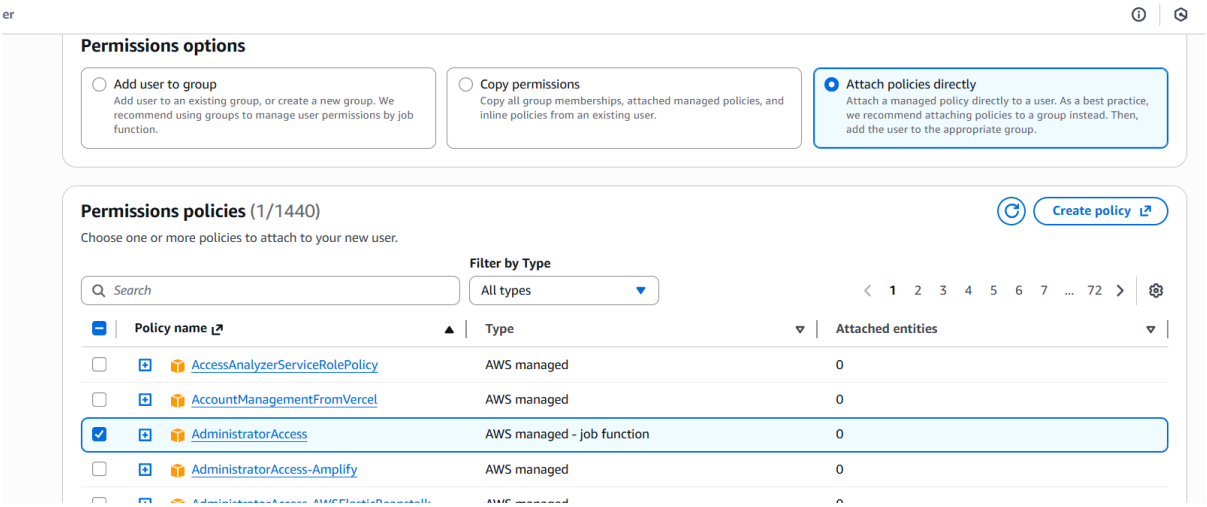
Creating a user:

Click on create user



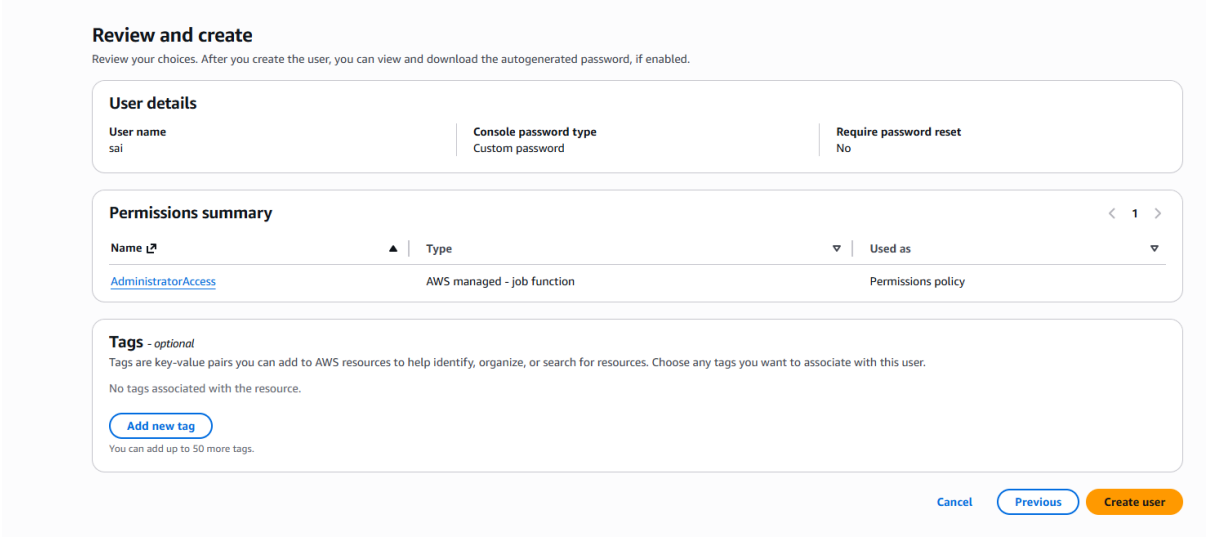


Click next



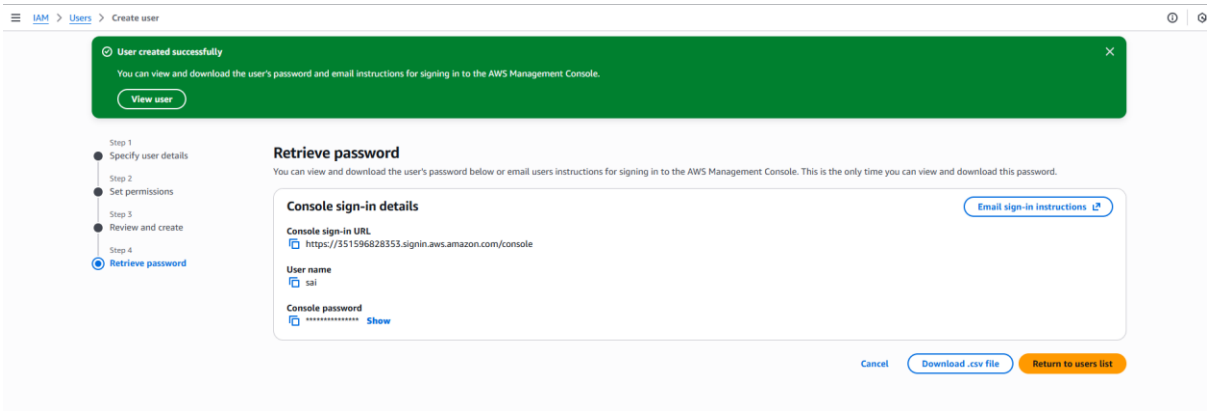
The above implies the user (sai) get all the admin access

Click on next



Click create user

U see the interface as shown:

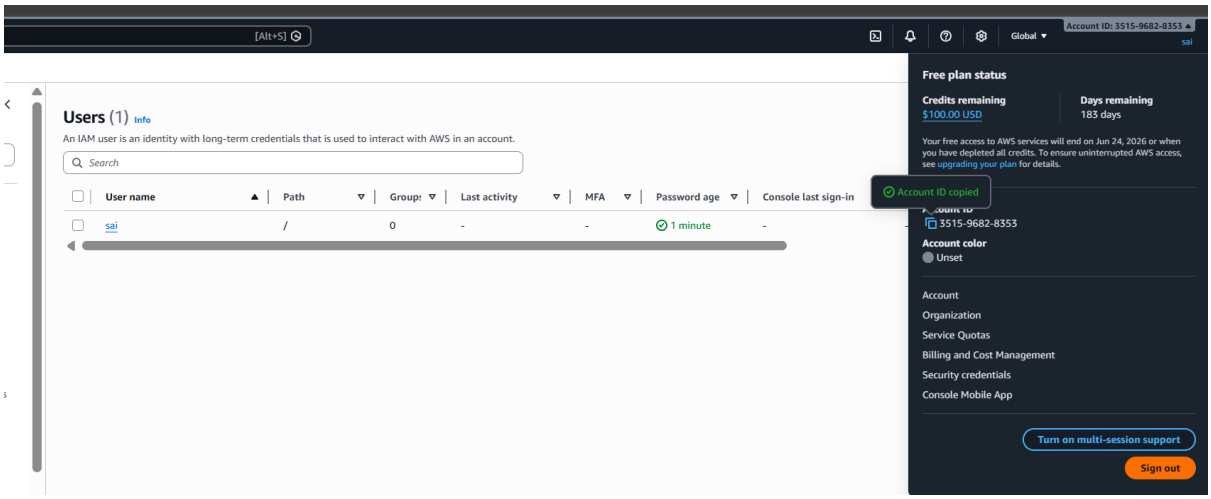


Now click return to users list

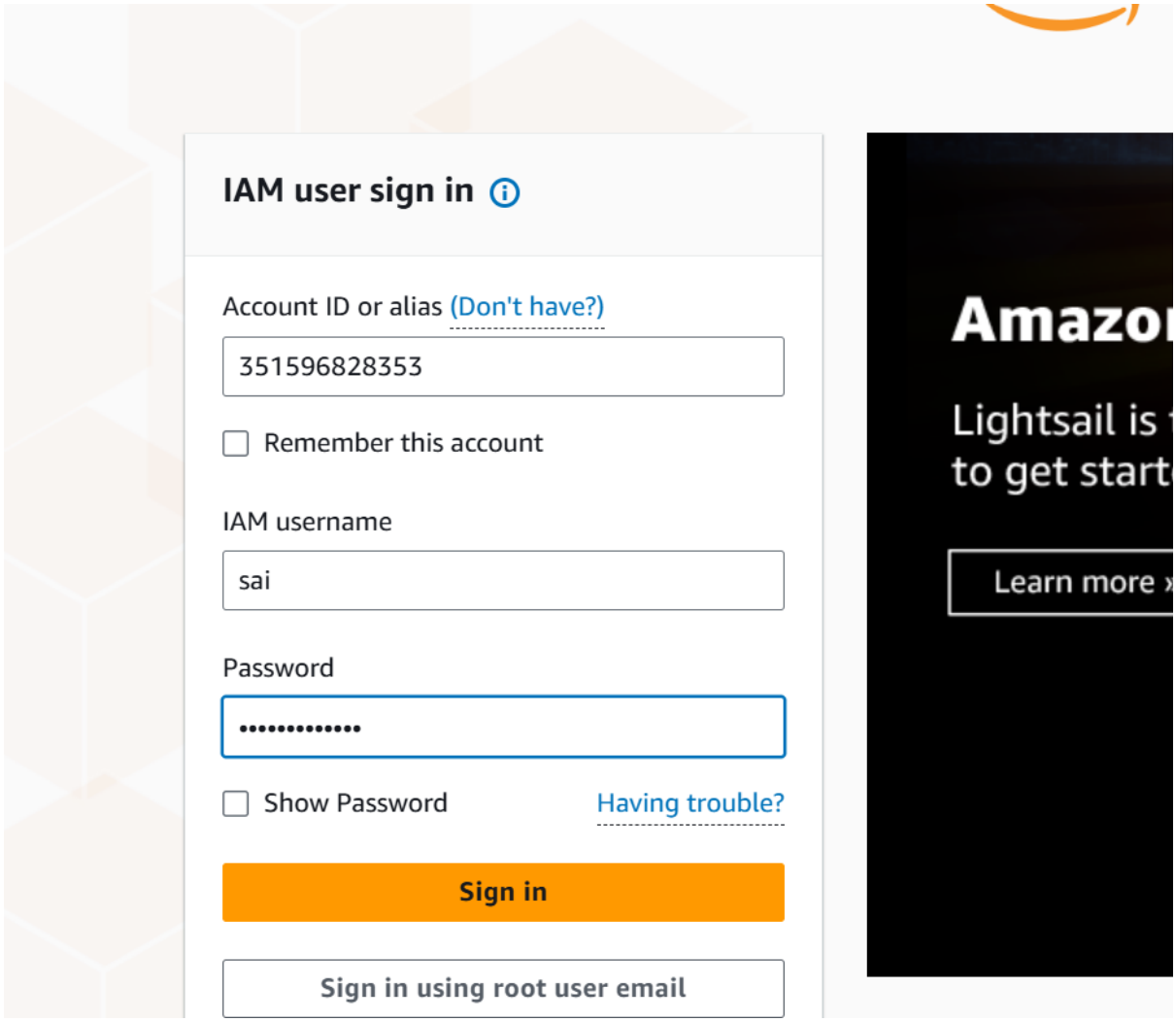
Log out and login as a IAM user

Before logging out, copy ur account ID: 351596828353

Click on sign out



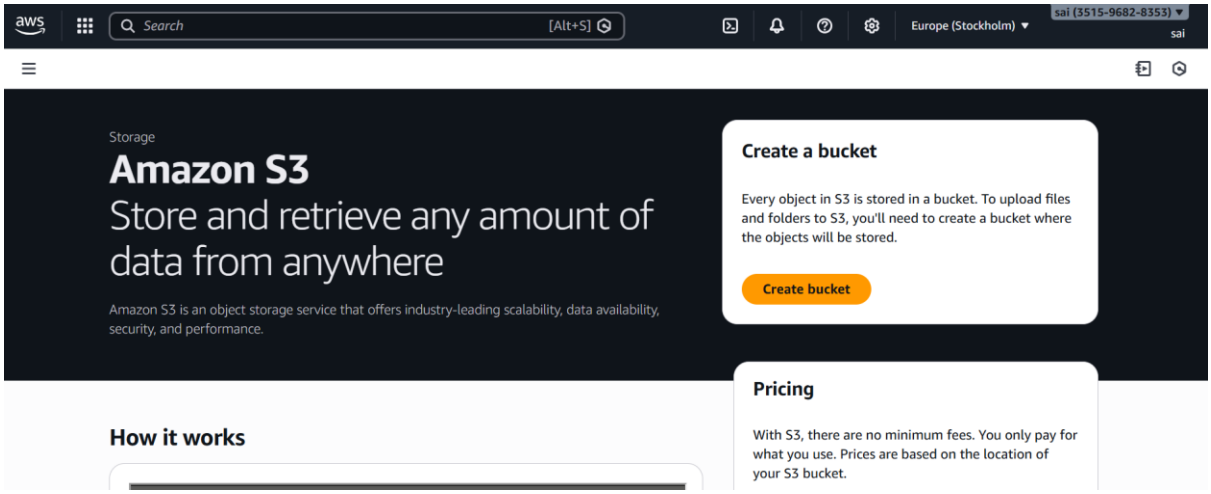
Logging as IAM user



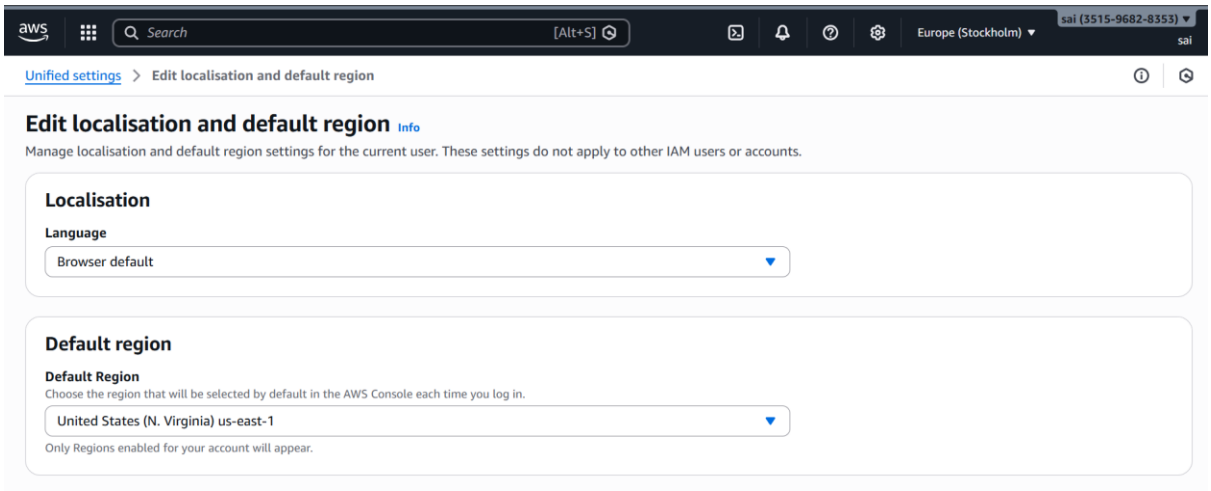
The interface looks like this and now click on s3 to create my bucket

S3 is a kind of cloud storage where u can store all kinds of files like structured, unstructured etc

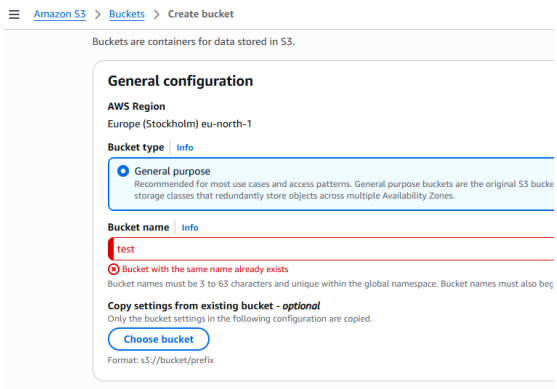
Let us create a bucket

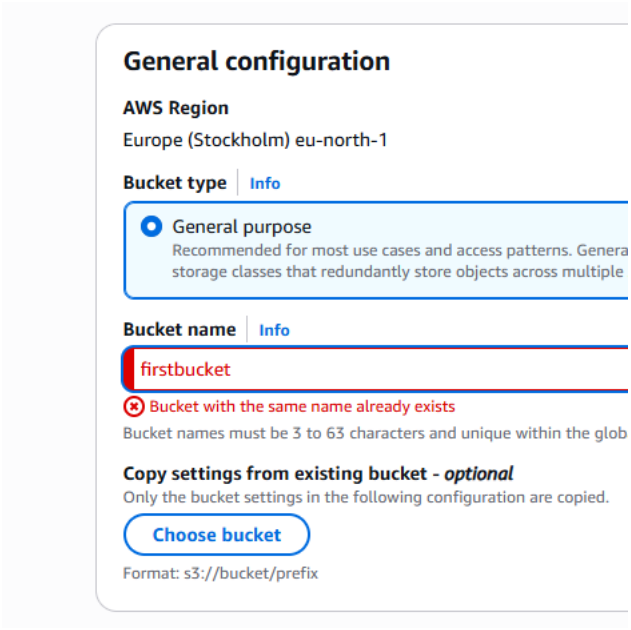


Before that changed the region as follows, click on setting button below beside the name

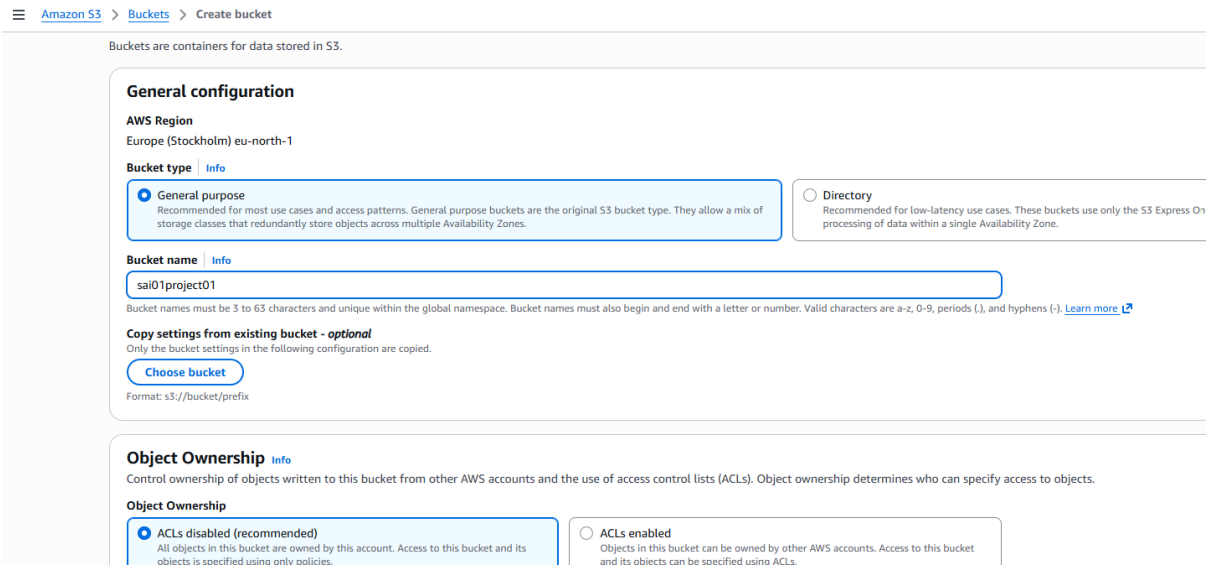


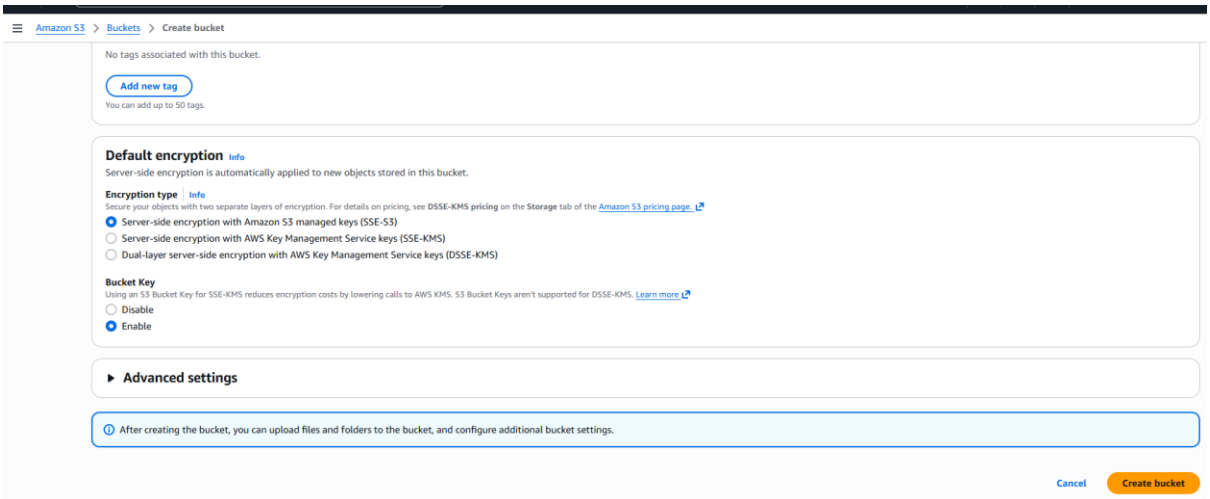
If u give the bucket name as test, as it is a name globally.so give a name which is unique globally



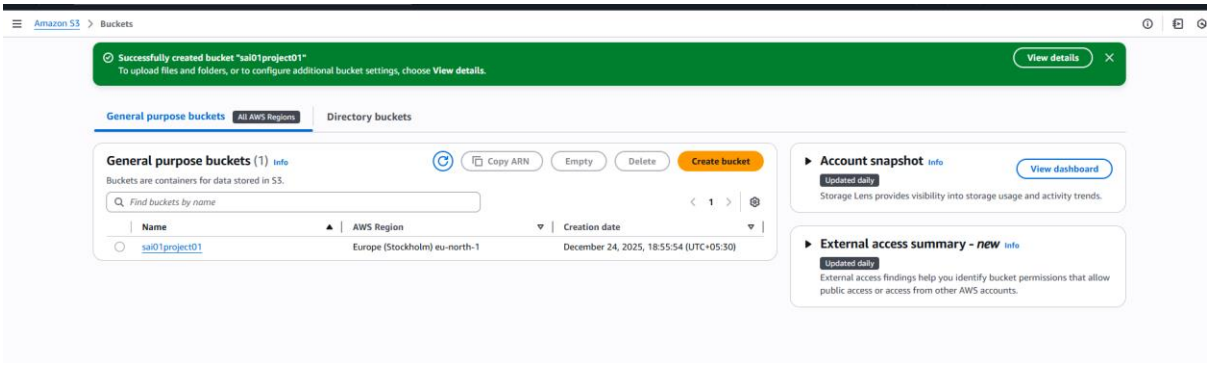


Keep all the default settings, I have given bucket name as ‘sai01project01’



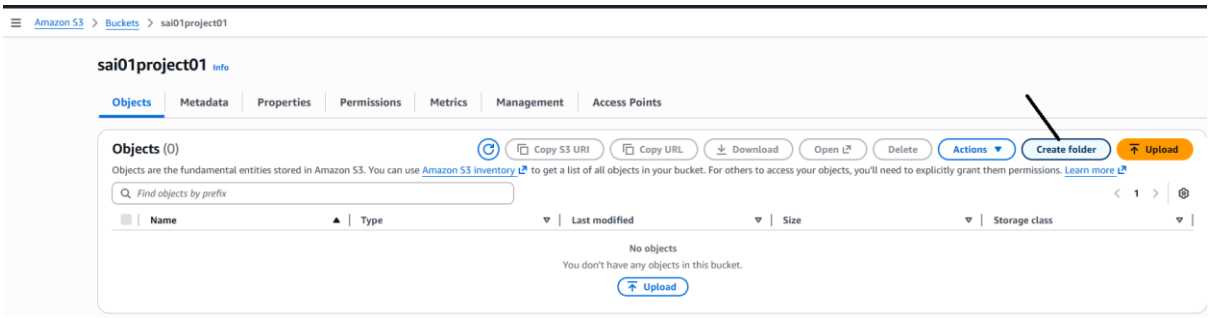


Click create bucket

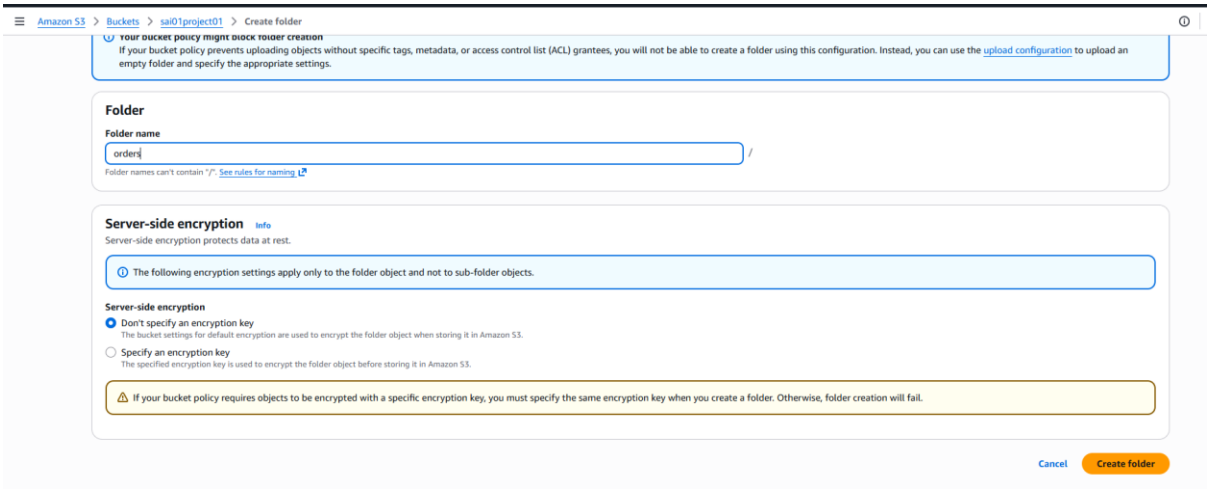


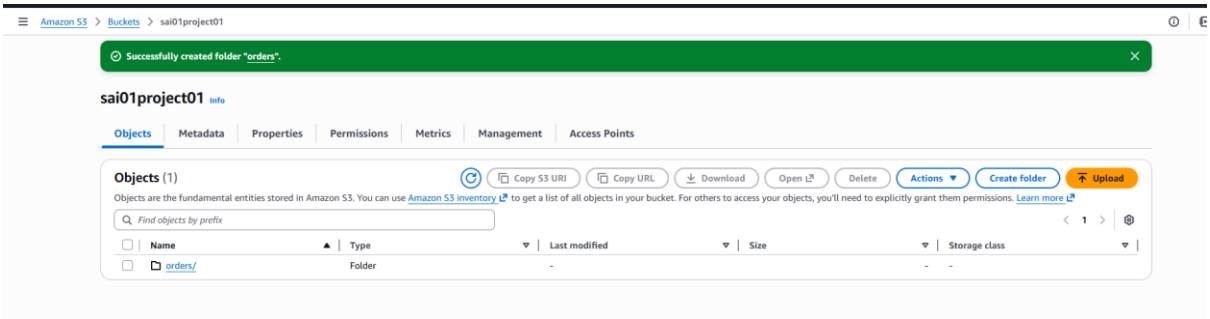
Click on it

Create a folder in it



Create folder as shown





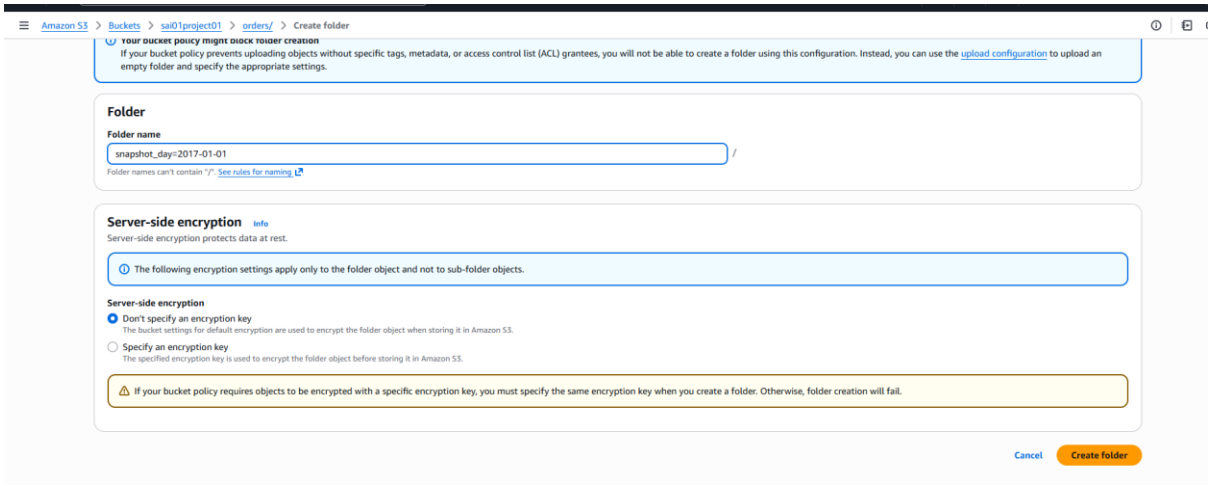
Lets download the Kaggle data
https://www.kaggle.com/datasets/vivek468/superstore-dataset-final?resource=download

Open the excel, filter ‘order date’ with jan 1st 2017, copy and paste in notepad, save

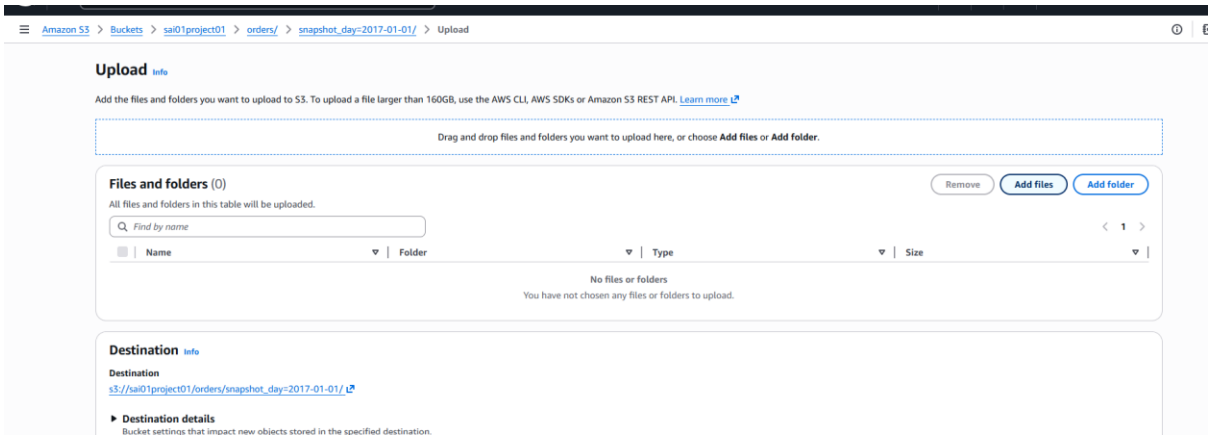
The data looks like this

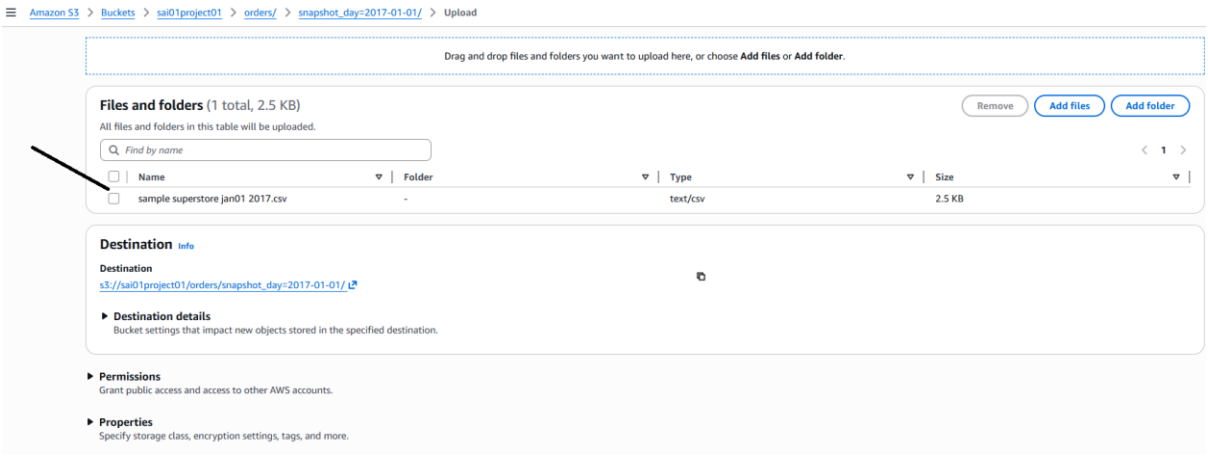
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	Row ID	Order ID	Order Date	Ship Date	Ship Mode	Customer	Customer	Segment	Country	City	State	Postal Coc	Region	Product ID	Category	Sub-Categ	Product Na	Sales
850	849	CA-2017-1	01-01-2017	########	Standard C	GA-14725	Guy Armst	Consumer	United Sta	Lorain	Ohio	44052	East	FUR-FU-1(	Furniture	Furnishing	Linden 10"	48.896
4011	4010	CA-2017-1	01-01-2017	########	Standard C	SC-20725	Steven Cai	Consumer	United Sta	Los Angele	California	90036	West	FUR-FU-1(	Furniture	Furnishing	Howard Mi	474.43
6684	6683	CA-2017-1	01-01-2017	########	First Class	DP-13390	Dennis Pai	Home Offi	United Sta	Franklin	Wisconsin	53132	Central	OFF-BI-10	Office Sup	Binders	Wilson Jon	3.6
8071	8070	CA-2017-1	01-01-2017	########	Standard C	JM-15250	Janet Mart	Consumer	United Sta	Huntsville	Texas	77340	Central	OFF-ST-10	Office Sup	Storage	SAFCO Bo	454.56
8072	8071	CA-2017-1	01-01-2017	########	Standard C	JM-15250	Janet Mart	Consumer	United Sta	Huntsville	Texas	77340	Central	FUR-FU-1(	Furniture	Furnishing	Tenex Car	141.42
8073	8072	CA-2017-1	01-01-2017	########	Standard C	JM-15250	Janet Mart	Consumer	United Sta	Huntsville	Texas	77340	Central	FUR-CH-1(	Furniture	Chairs	Office Star	310.744
8074	8073	CA-2017-1	01-01-2017	########	Standard C	JM-15250	Janet Mart	Consumer	United Sta	Huntsville	Texas	77340	Central	OFF-AR-1(	Office Sup	Art	Fluorescer	12.736
8075	8074	CA-2017-1	01-01-2017	########	Standard C	JM-15250	Janet Mart	Consumer	United Sta	Huntsville	Texas	77340	Central	OFF-BI-10	Office Sup	Binders	GBC Instar	6.47
8076	8075	CA-2017-1	01-01-2017	########	Standard C	JM-15250	Janet Mart	Consumer	United Sta	Huntsville	Texas	77340	Central	OFF-BI-10	Office Sup	Binders	Pressboar	13.748
8077	8076	CA-2017-1	01-01-2017	########	Standard C	JM-15250	Janet Mart	Consumer	United Sta	Huntsville	Texas	77340	Central	OFF-AP-1(	Office Sup	Appliance	Fellowes S	15.224
9996																		
9997																		

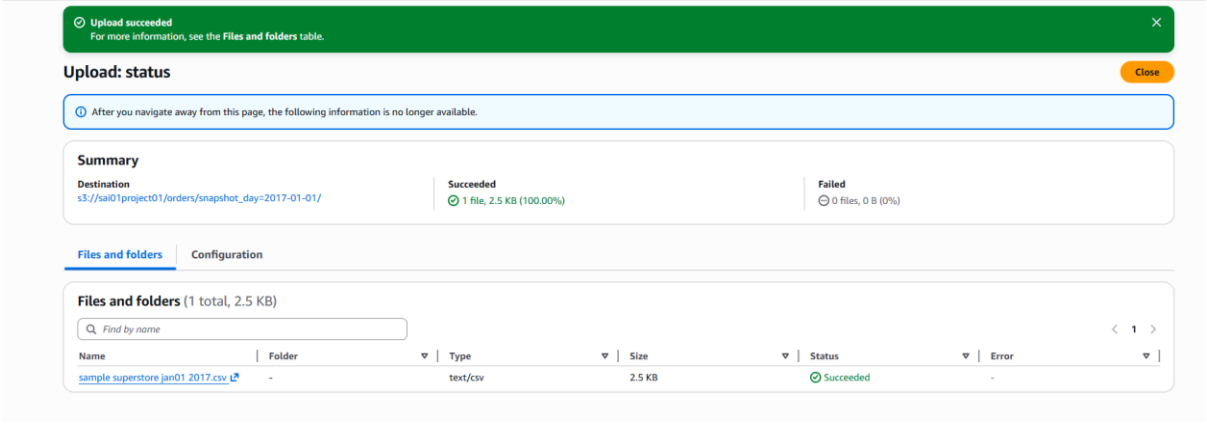
Again creating a folder in orders



When we create AWS crawler, when we create the structure of the folder with the name like this, there is a partition with all of this, we can discuss later





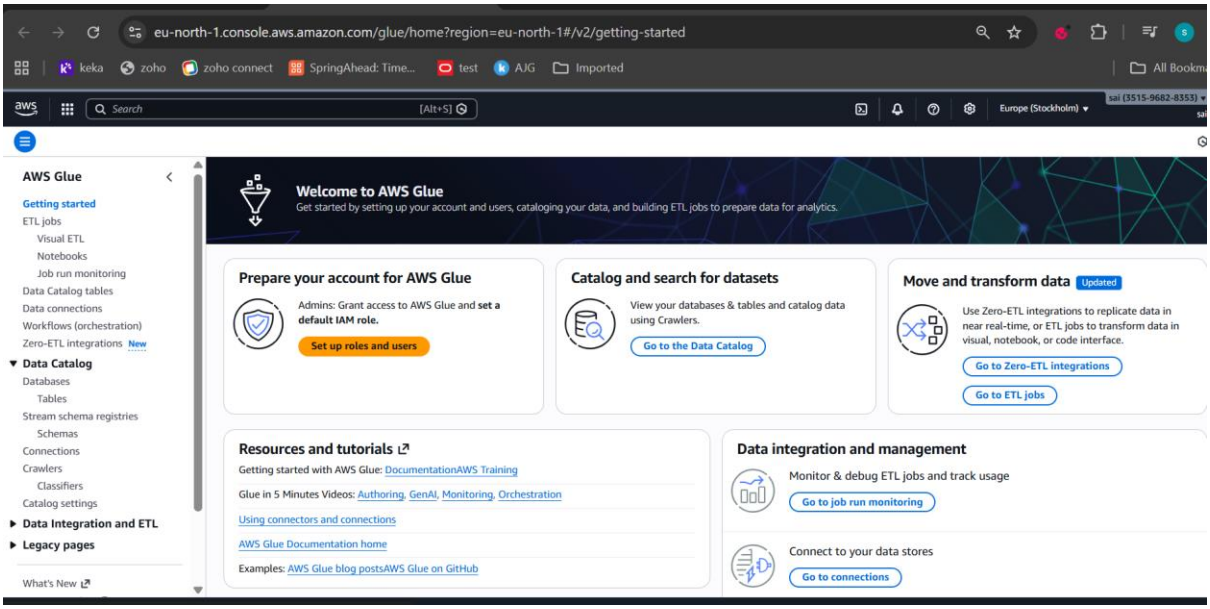


Now let us create AWS crawler

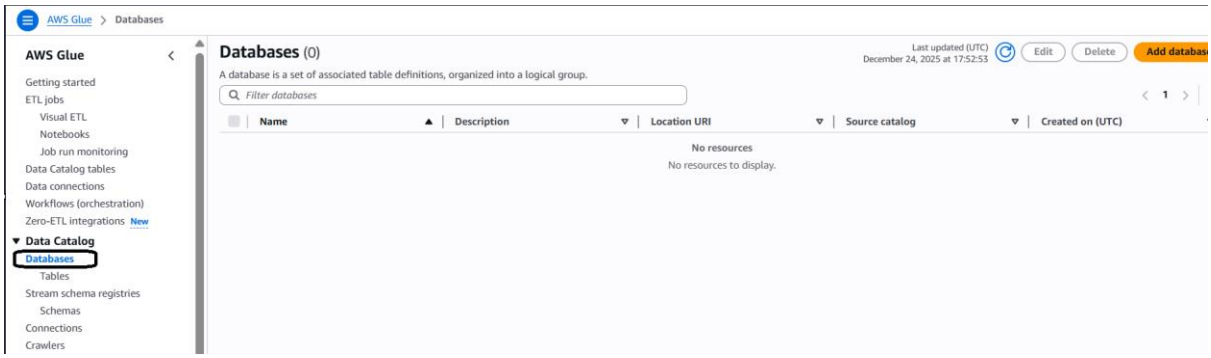
Search glue

Glue is an ETL tool and u can run also crawler and create a data catalog using glue

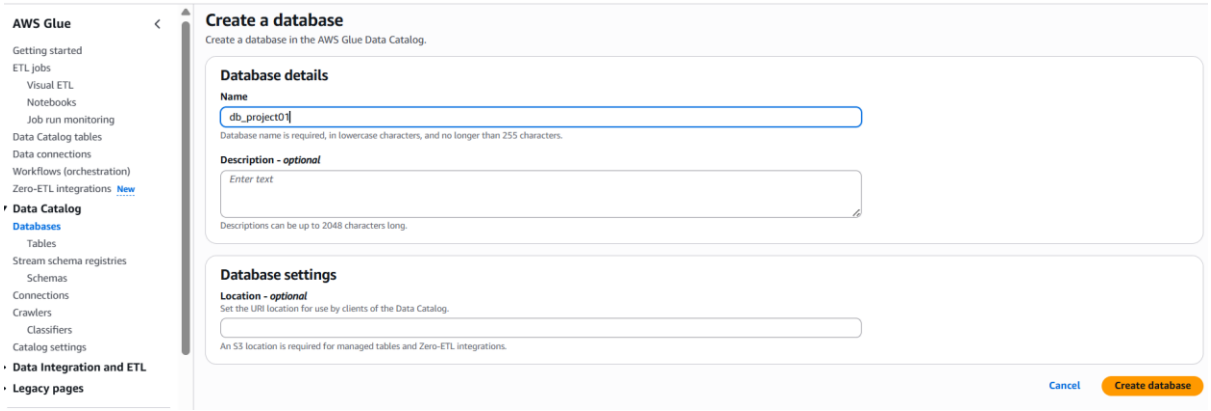
Here is the GLUE interface



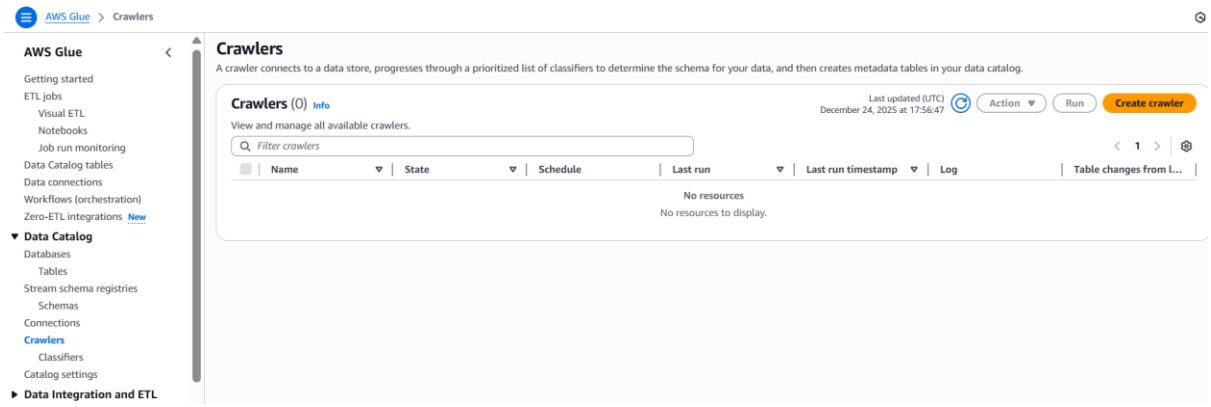
We are going into crawler and data cateloging



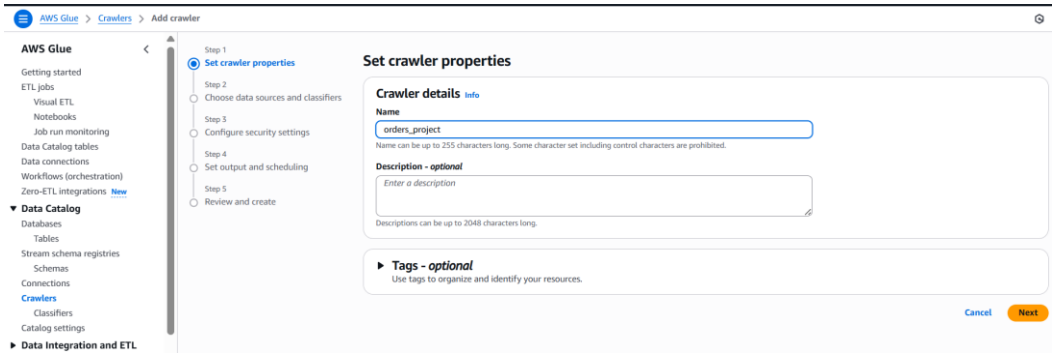
Click on add data base ,give name and click create database



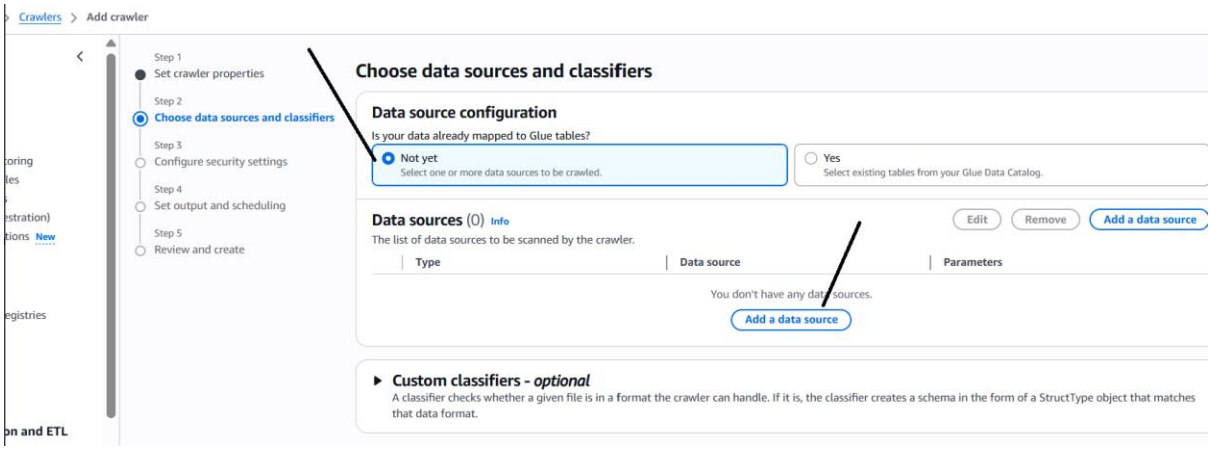
Click on crawlers, create crawler

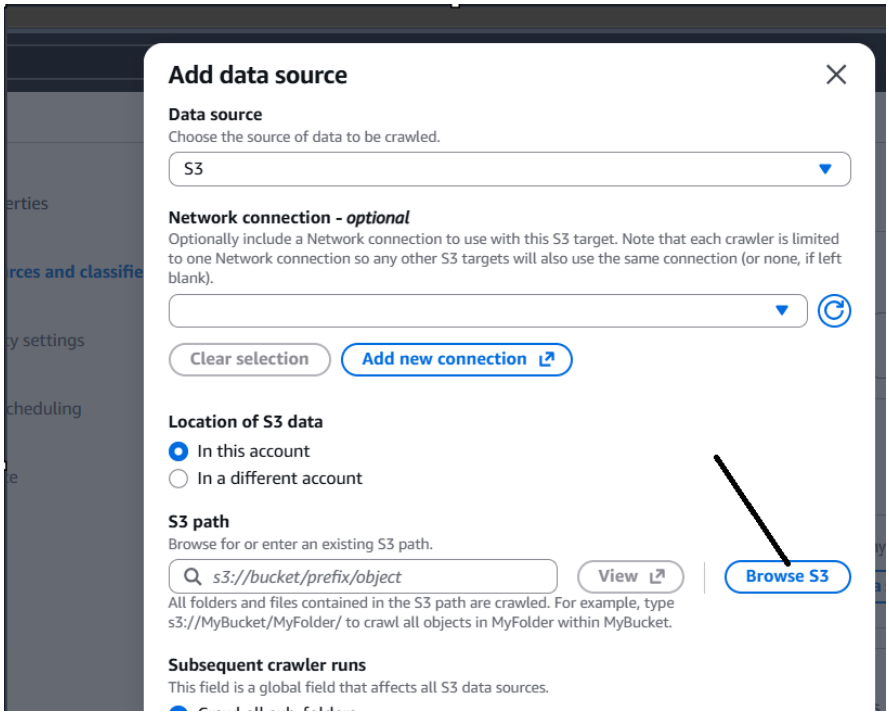


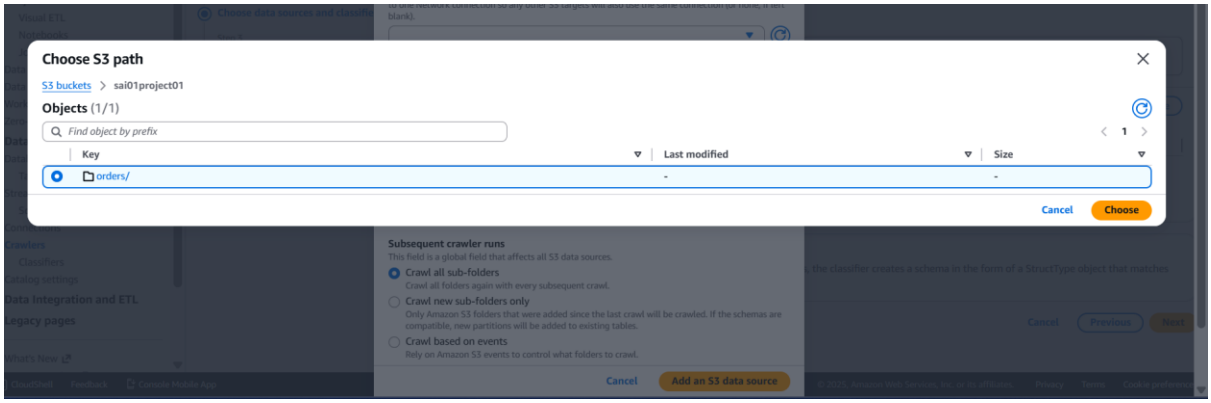
Give name and next

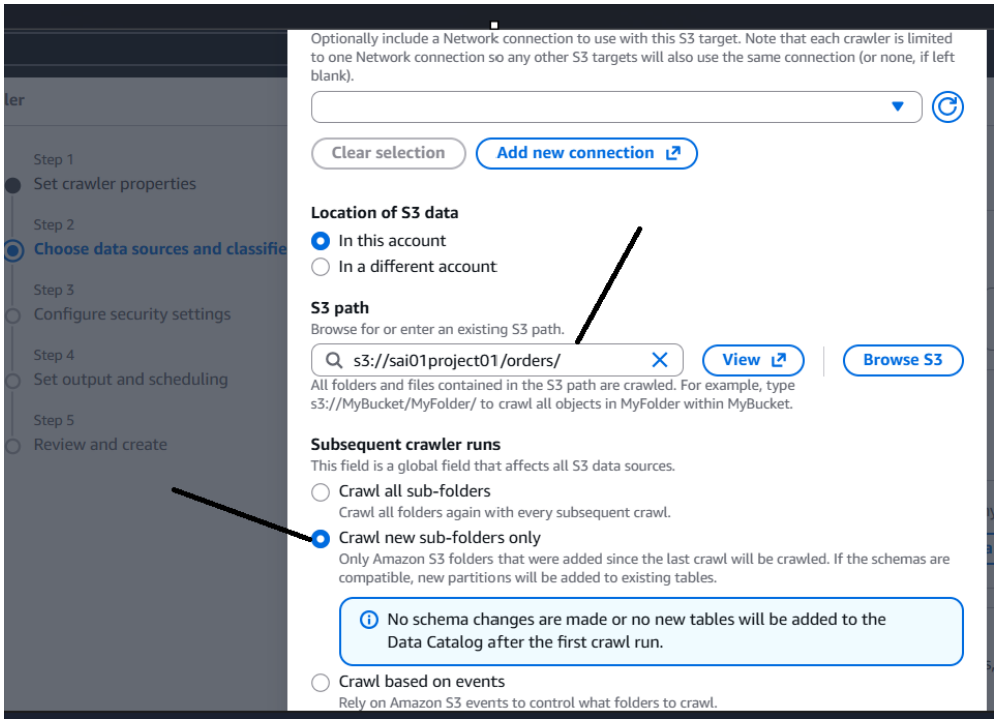


Click Add datasource

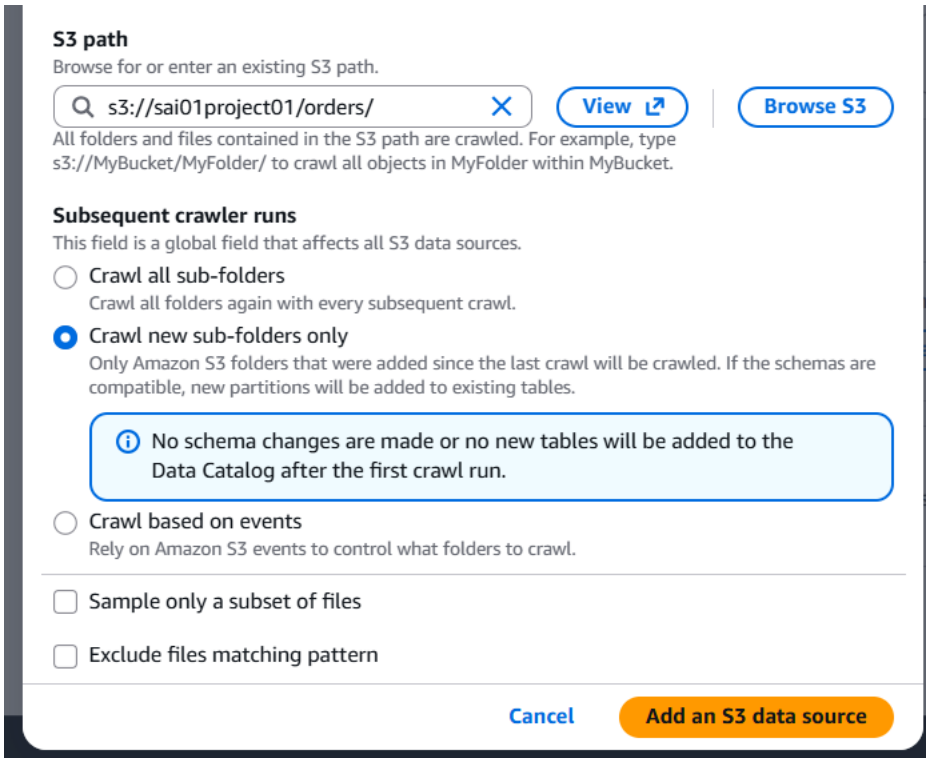






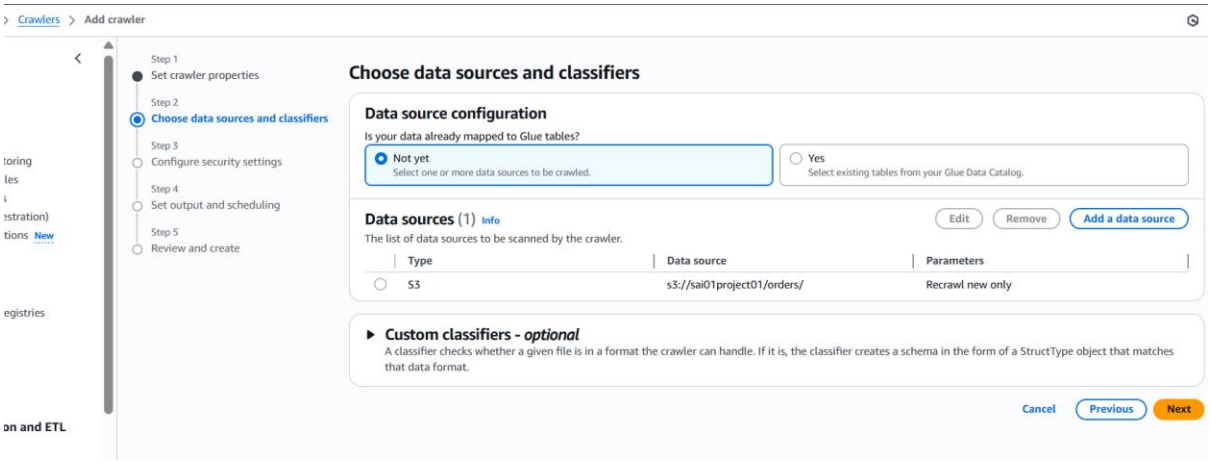


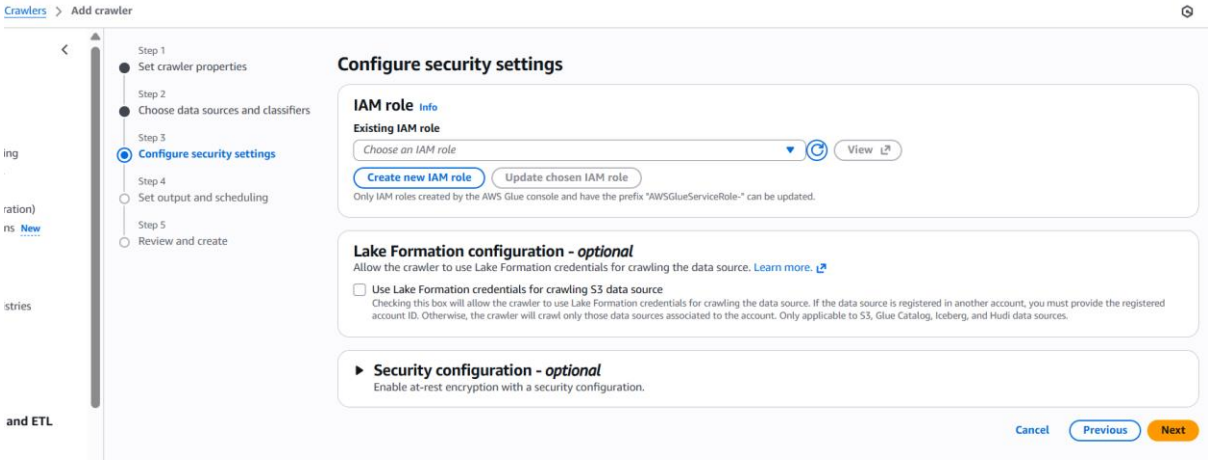
Click add an s3 datasource



The above meaning is the folders which were crawled will not be crawled after addition of new files

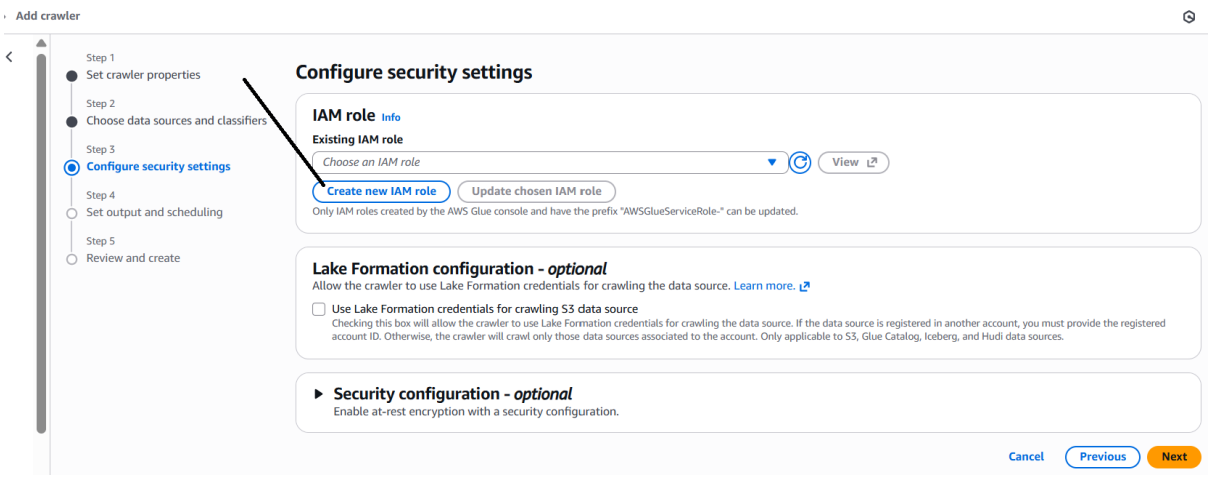
Click next



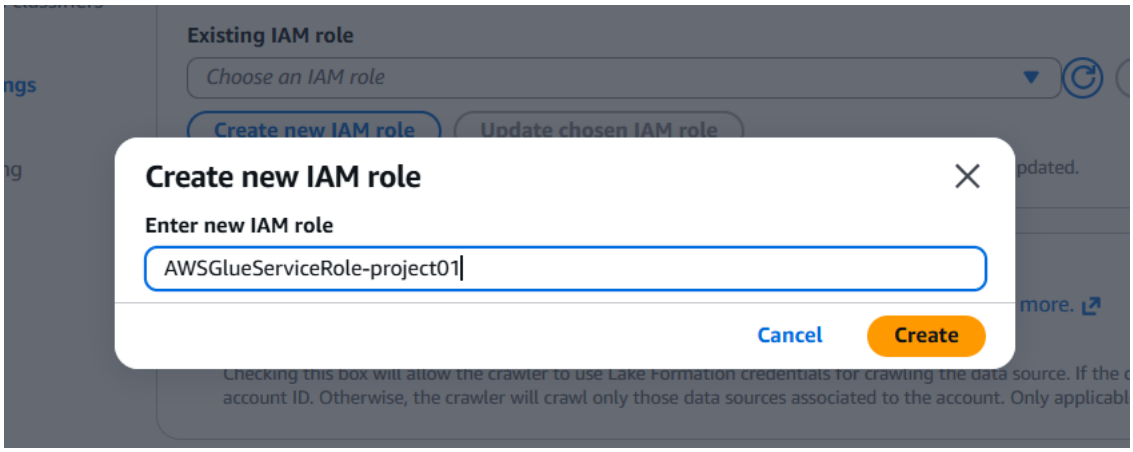


We have created IAM user and user logged in and using AWS services. similarly for AWS services to work they need some access . so, IAM role is used to give different services access.

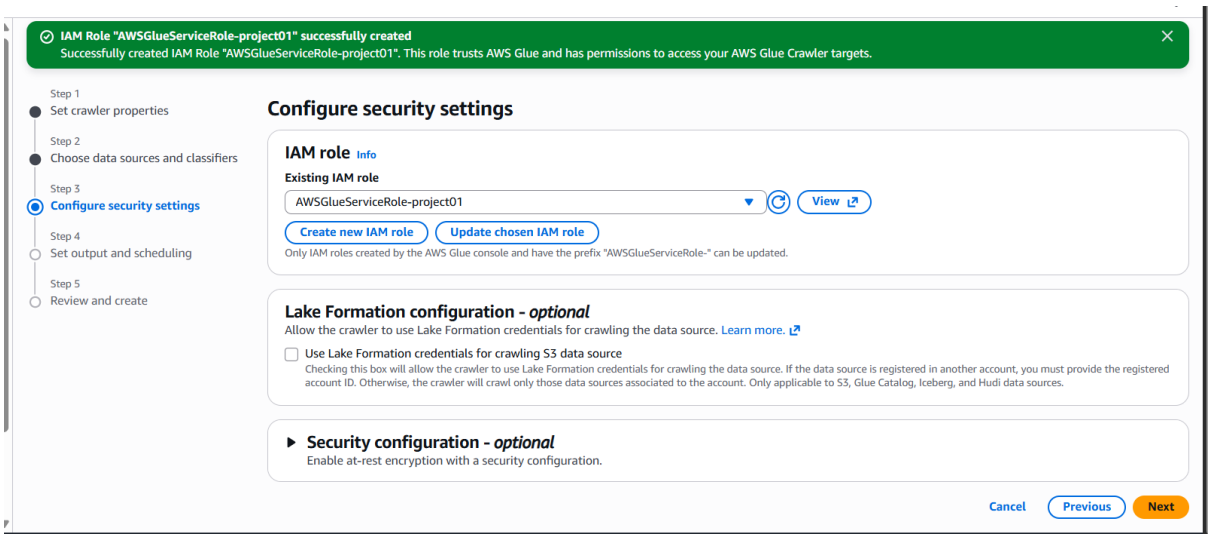
As I Want to run the AWS crawler job, I have to use an IAM role for that using which role I want to run the services and to that role I have to give all the access which it needs.

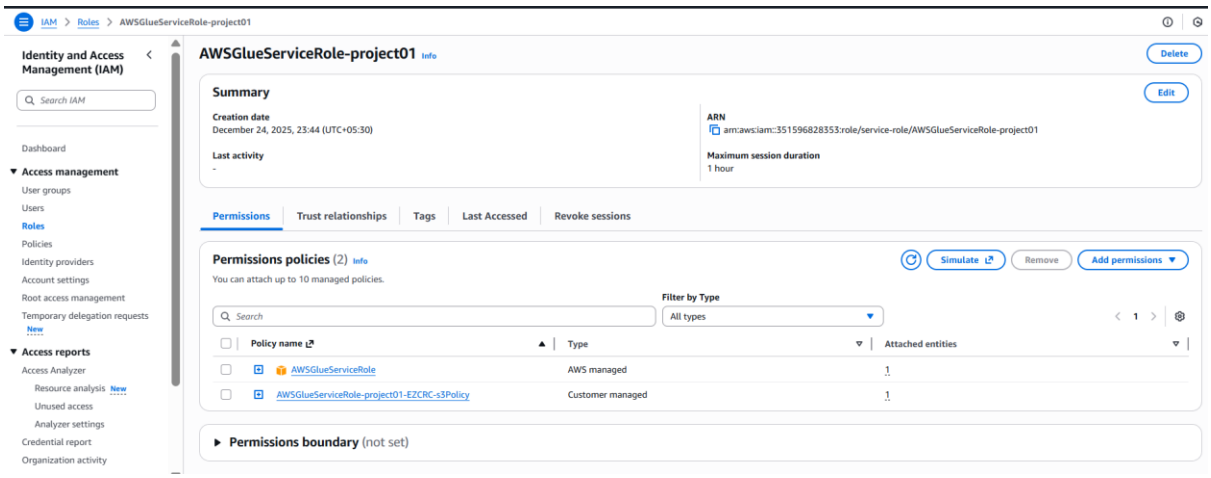


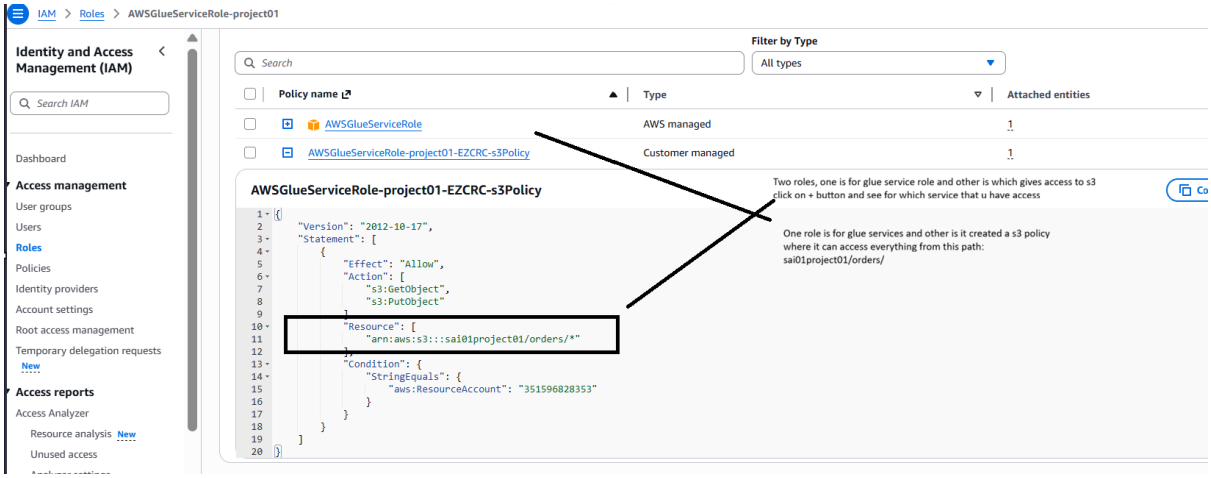
Project01 is the name and AWSGlueServiceRole is by default available



We had created IAM role as shown, click on view to see

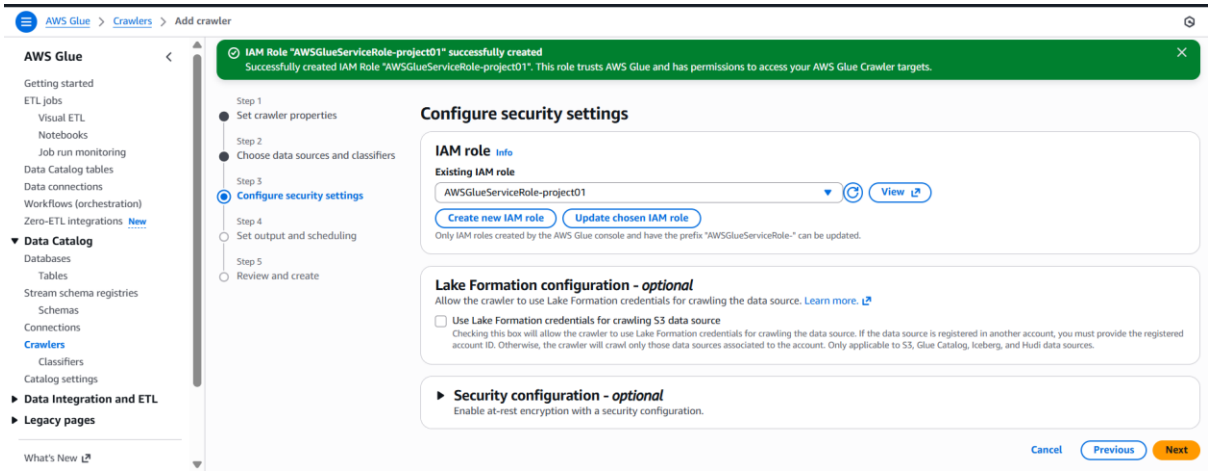


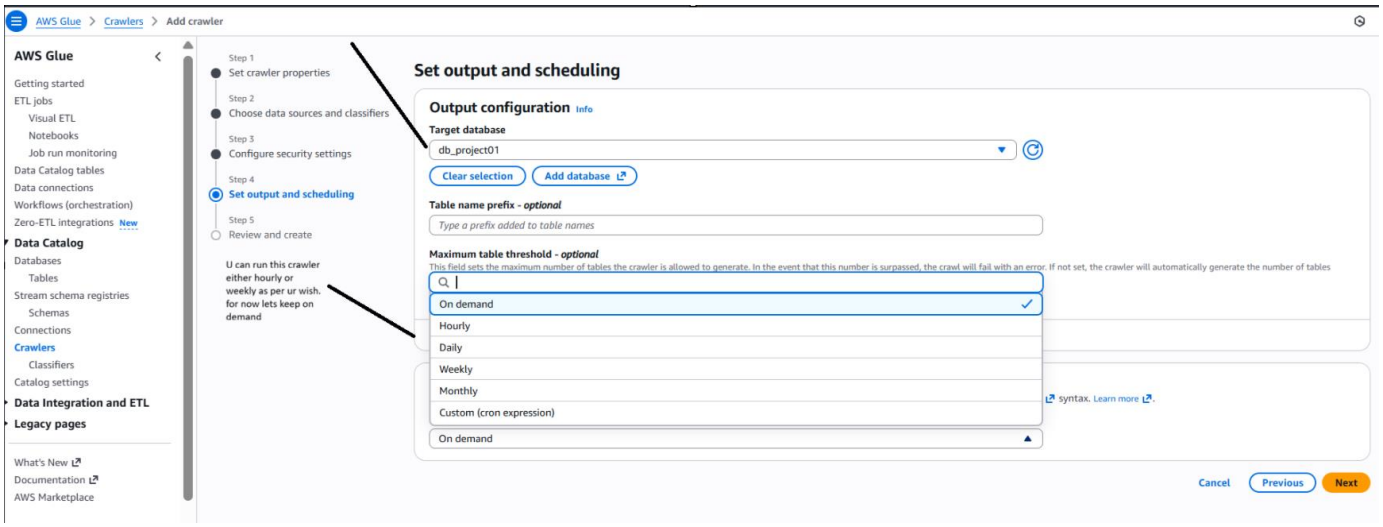




As we have given data source as S3 path, it has automatically taken the access to that s3 path by creating a policy

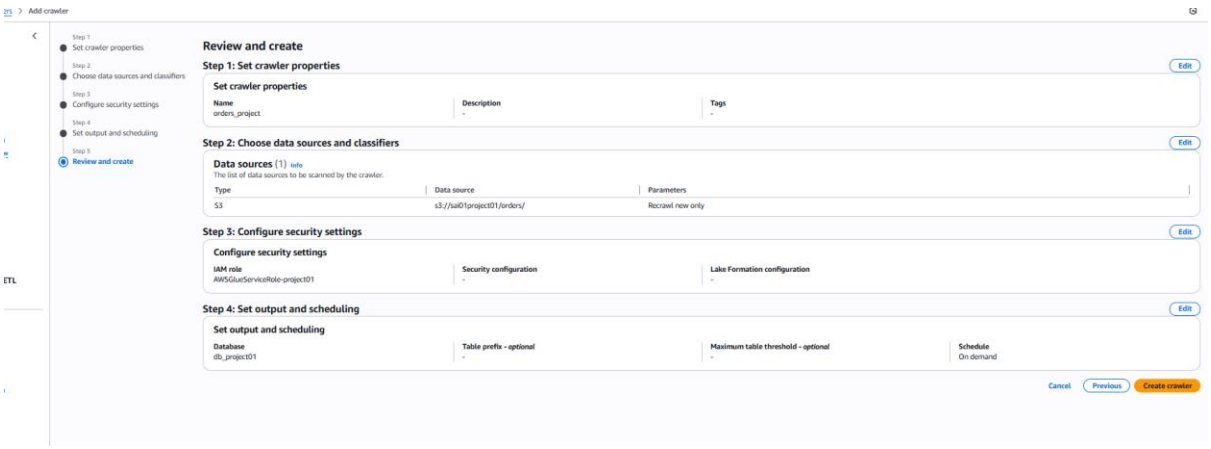
Click next



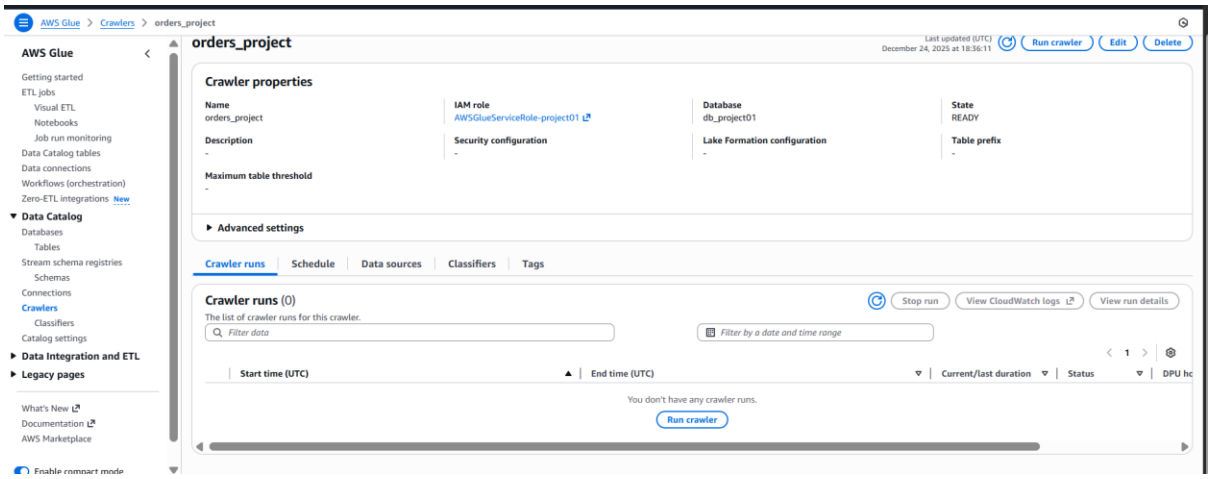


Click on next

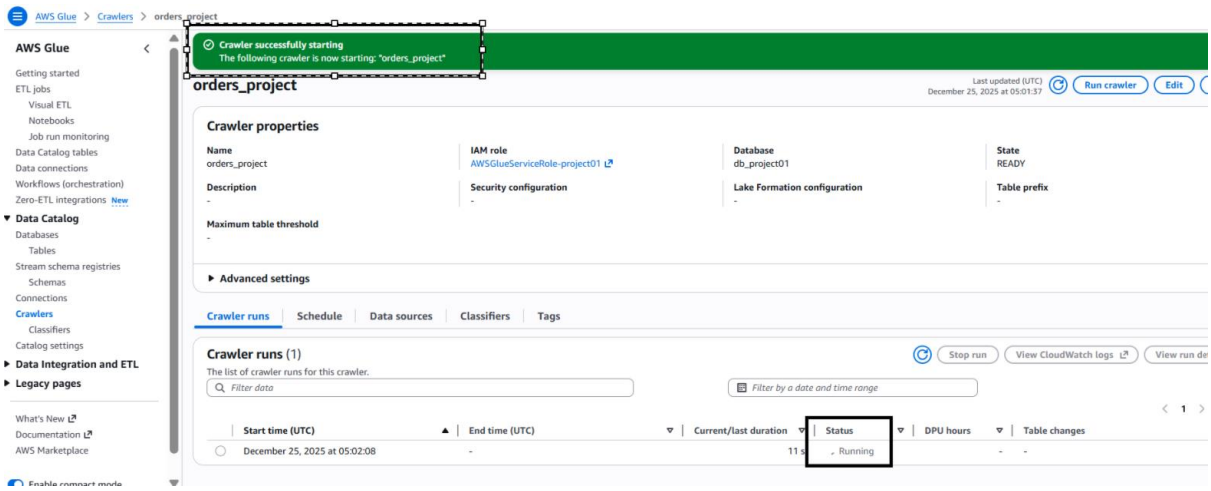
Create crawler



Here is the interface



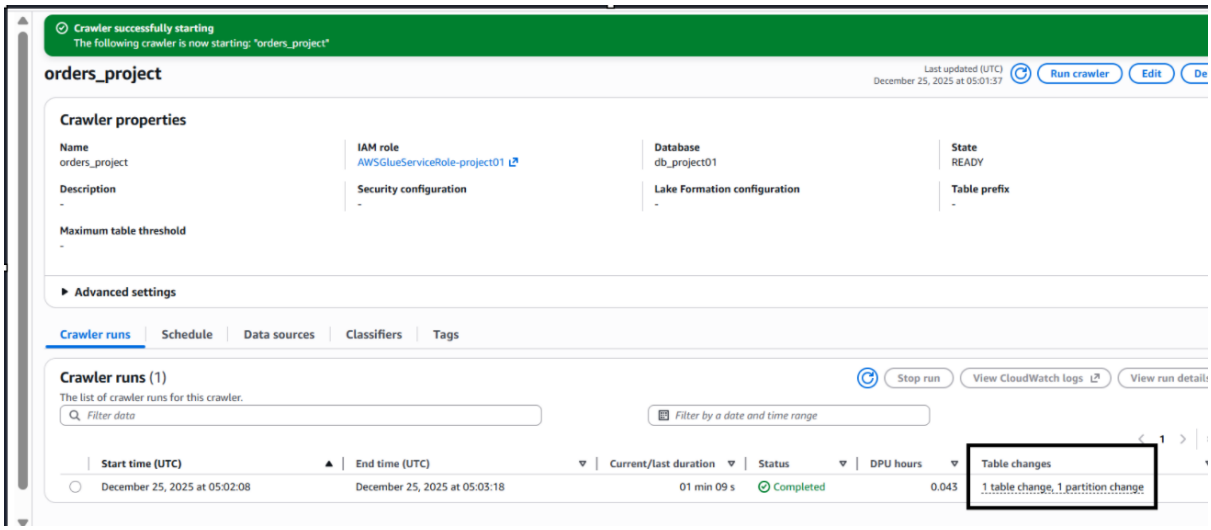
Click on run crawler

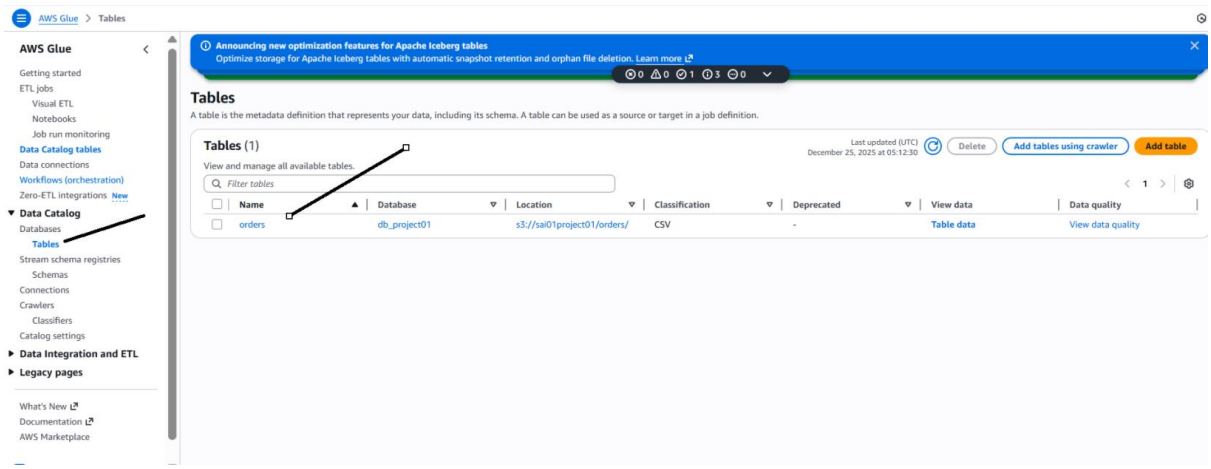


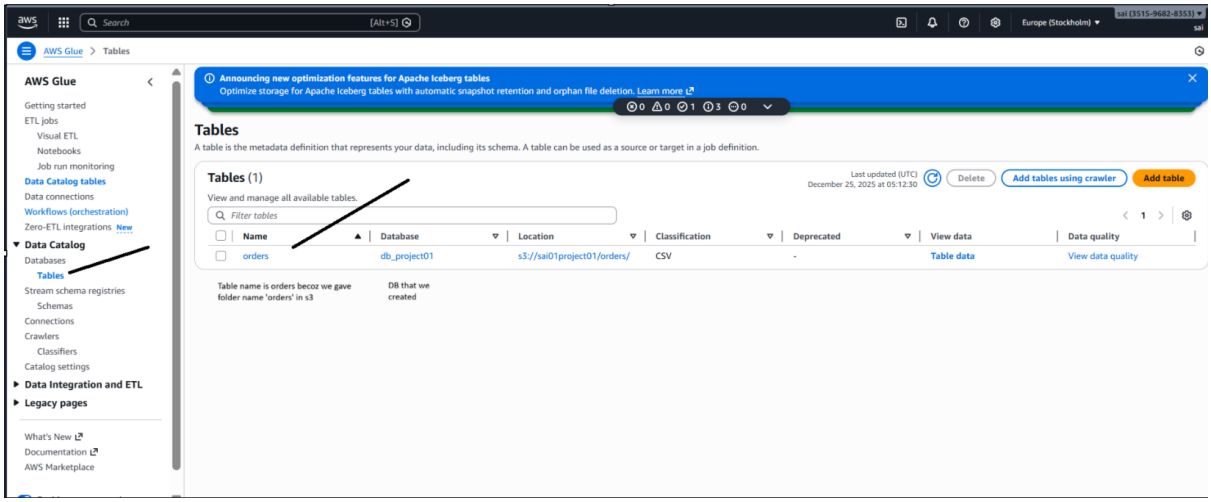
Working of crawler…

As our data is in s3 which is a file, if u want to query this data u need to have information about data which is meta data like what are the columns, datatypes etc.Then only I can query the data using sql.

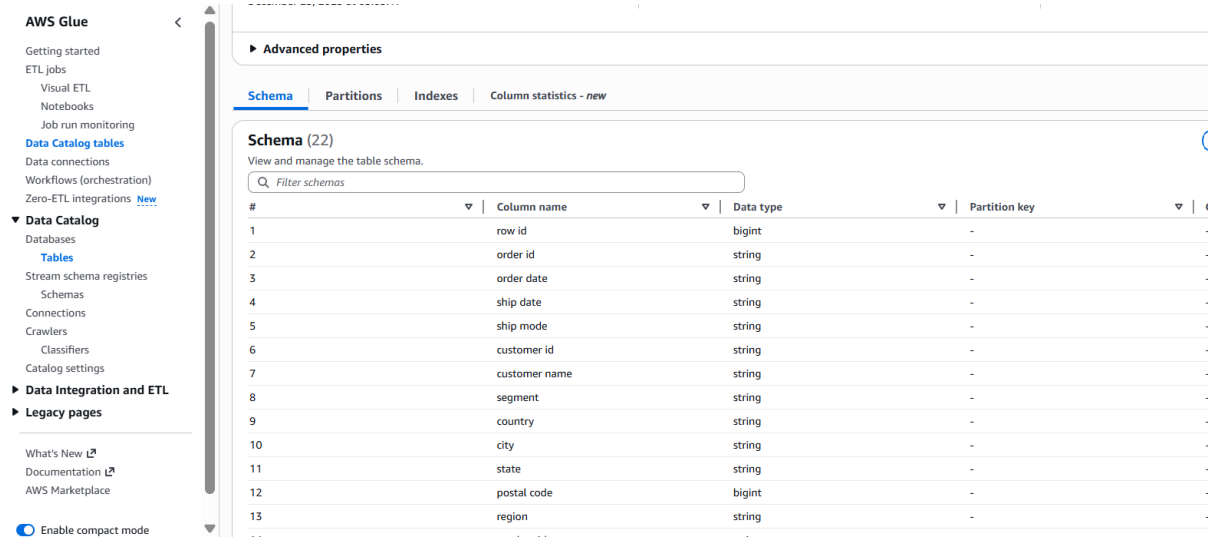
So this crawler crawls over the s3 file, identify what are the columns,delimiters etc in the file internally by itself and u will be able to see table but it doesn't store any data in it, the data is in s3 only, it is just creating meta data

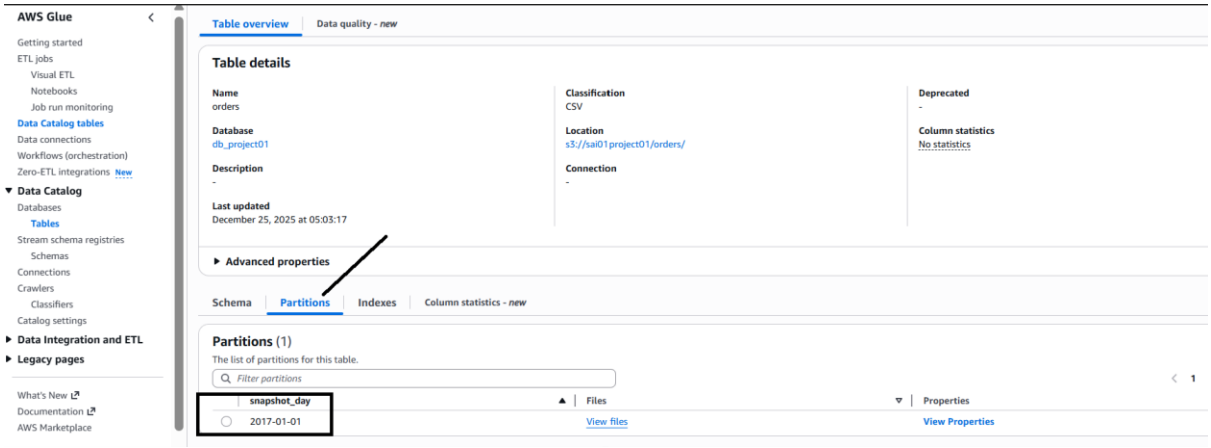






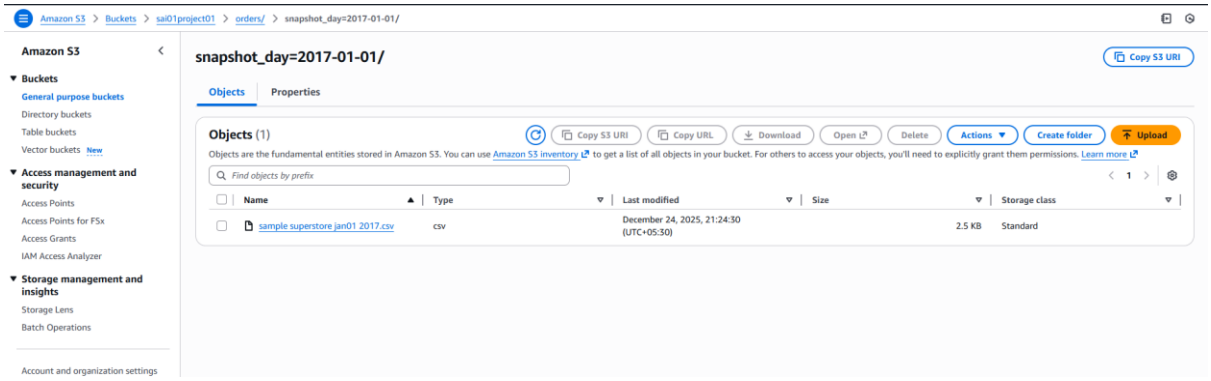
Click on the orders table, u see the meta data as shown





The crawler created a partition already, if I created one more folder and run the crawler again it will create another partition, these partitions will help in querying in athena

So now, if u click on view files, u see one file



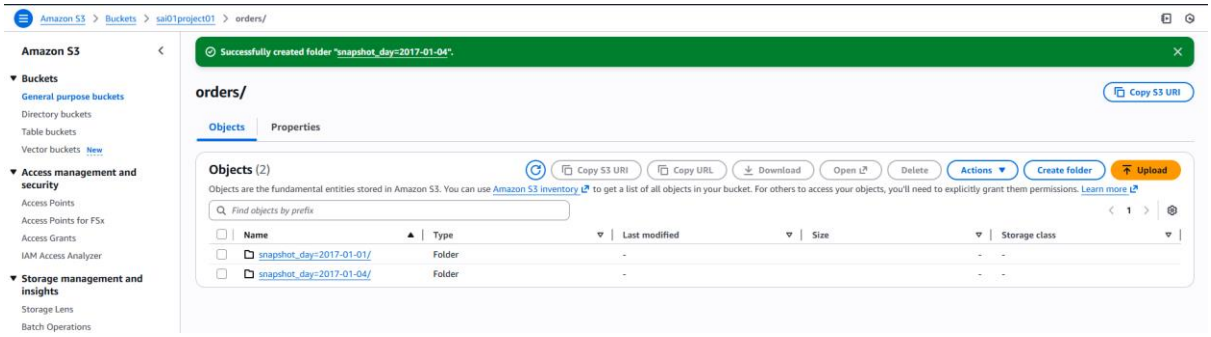
So u can click on every partition and see which s3 files are available

Let us create another folder in s3

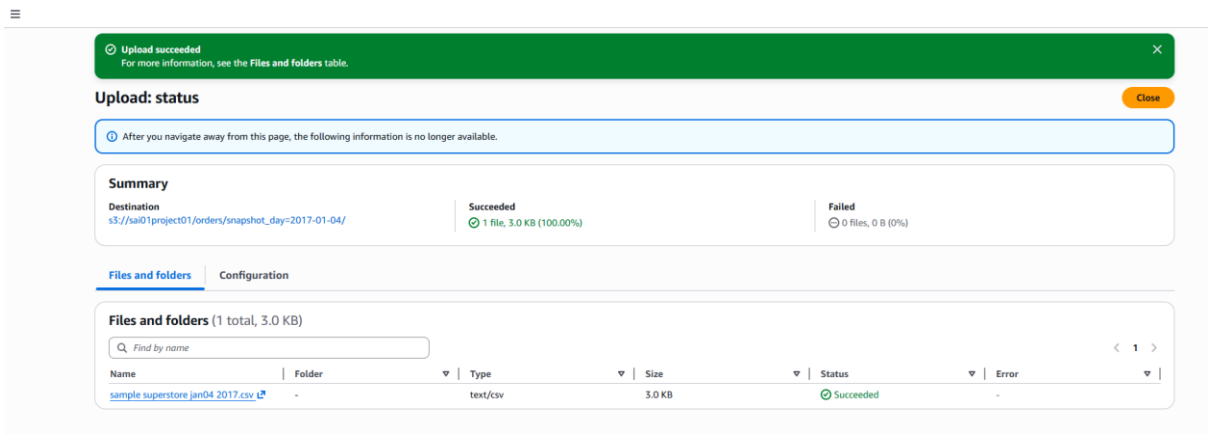
Filtered to 4rth jan, copy in notepad, save it

	A	B	C	D	E	F	G	H	I	J	K	L
1	Row ID	Order ID	Order Da	Ship Date	Ship Mode	Customer	Customer	Segment	Country	City	State	Postal Coc
2403	2402	CA-2017-145877	04-01-2017	########	Second Cl	AS-10090	Adam Shill	Consumer	United Sta	Springfield	Missouri	65807
2404	2403	CA-2017-145877	04-01-2017	########	Second Cl	AS-10090	Adam Shill	Consumer	United Sta	Springfield	Missouri	65807
2618	2617	CA-2017-147942	04-01-2017	########	Standard C	MS-17365	Maribeth S	Consumer	United Sta	San Franci	California	94110
2619	2618	CA-2017-147942	04-01-2017	########	Standard C	MS-17365	Maribeth S	Consumer	United Sta	San Franci	California	94110
4219	4218	CA-2017-149881	04-01-2017	########	First Class	NC-18535	Nick Creb	Corporate	United Sta	San Franci	California	94110
4220	4219	CA-2017-149881	04-01-2017	########	First Class	NC-18535	Nick Creb	Corporate	United Sta	San Franci	California	94110
4330	4329	CA-2017-118360	04-01-2017	########	Standard C	JC-15775	John Caste	Consumer	United Sta	New York C	New York	10011
5489	5488	CA-2017-134495	04-01-2017	########	Second Cl	BF-11020	Barry Fran	Corporate	United Sta	Jacksonvil	Florida	32216
5490	5489	CA-2017-134495	04-01-2017	########	Second Cl	BF-11020	Barry Fran	Corporate	United Sta	Jacksonvil	Florida	32216
5491	5490	CA-2017-134495	04-01-2017	########	Second Cl	BF-11020	Barry Fran	Corporate	United Sta	Jacksonvil	Florida	32216
5492	5491	CA-2017-134495	04-01-2017	########	Second Cl	BF-11020	Barry Fran	Corporate	United Sta	Jacksonvil	Florida	32216
6009	6008	CA-2017-129028	04-01-2017	########	First Class	GB-14530	George Be	Corporate	United Sta	Florence	South Car	29501
9996												

Created a folder in s3 with snapshot_day=2017-01-04



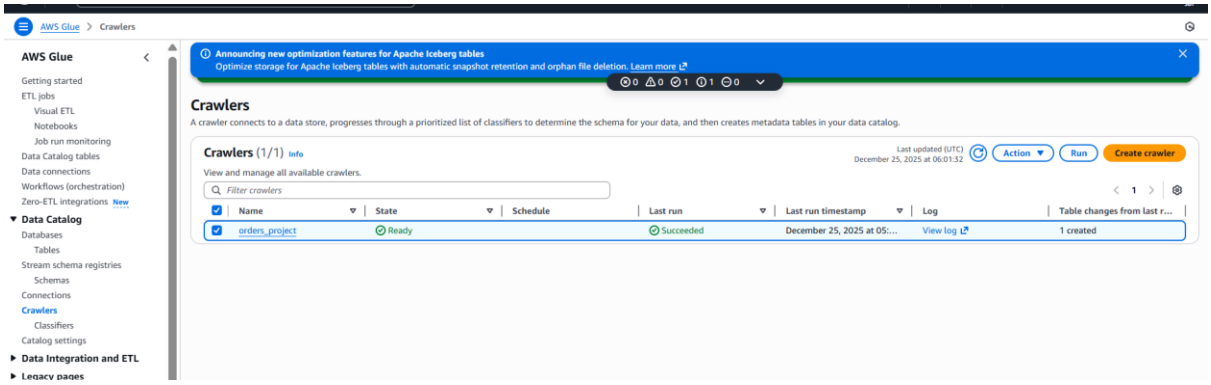
I uploaded my csv file in the folder



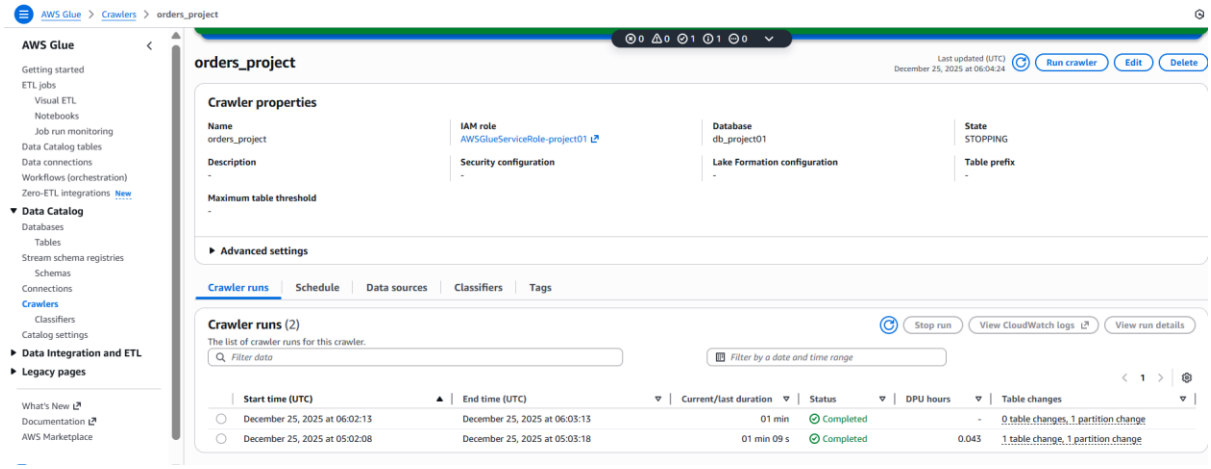
Note: u can add multiple files in it, np

Now let us run the crawler in glue as shown

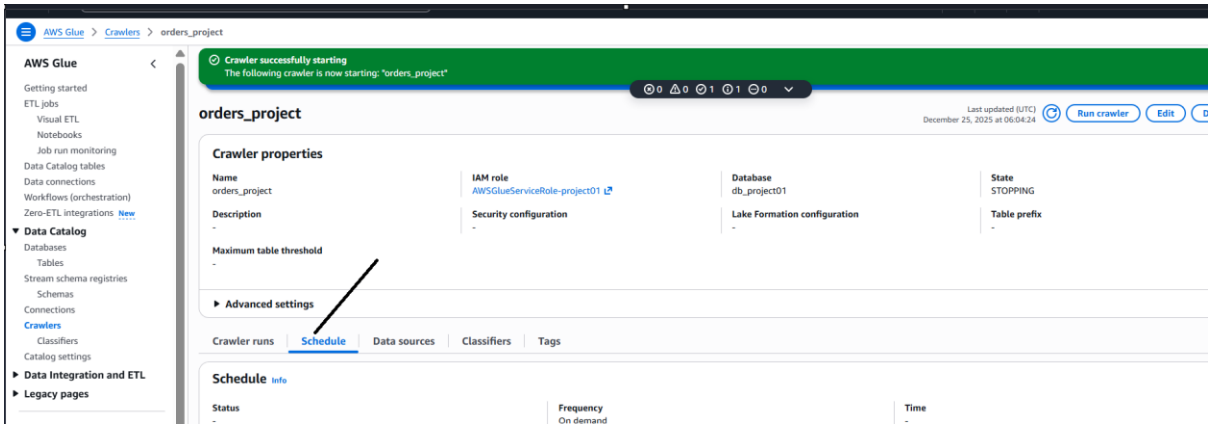
Click run



If u click on the crawler, u see the status and shows u all the run

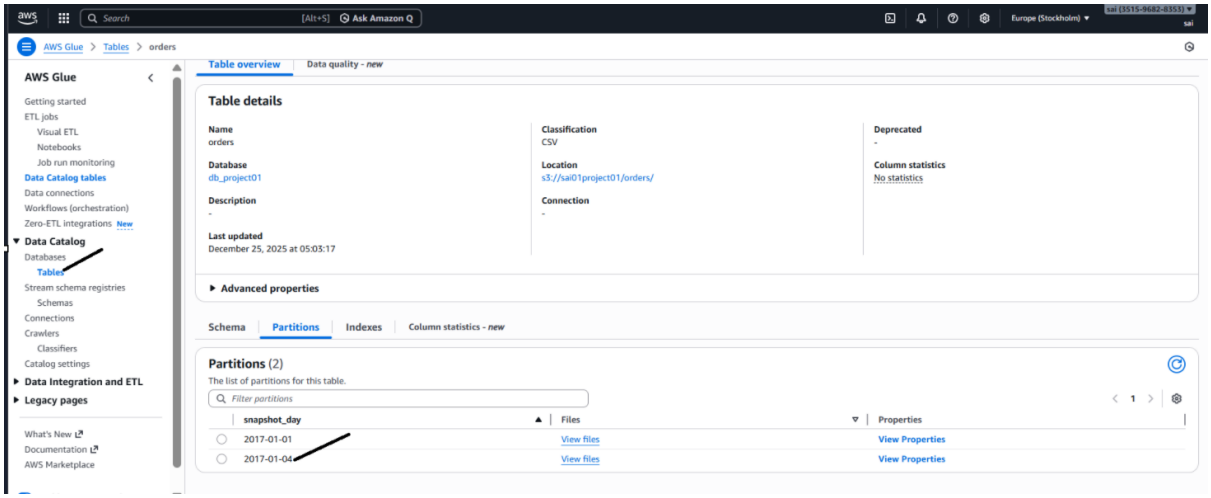


Now we are running on demand but in production u set a schedule,as ur files are coming daily overnight. Next day if u schedule at 5 am what ever the new folders had created it will crawl



While creating crawler, we had chosen crawl only new subfolders instead of crawl all folders

Now let us go to tables, u see another partition added as shown



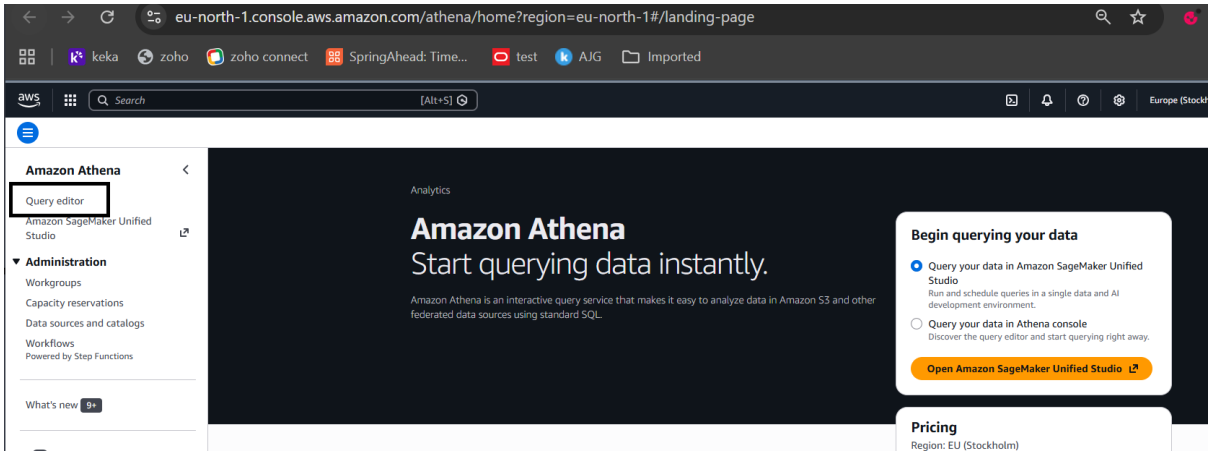
Lets say if my s3 folder is not given in the format snapshot like may be with only date, now u see it creates partition name with some random name eg: partition key

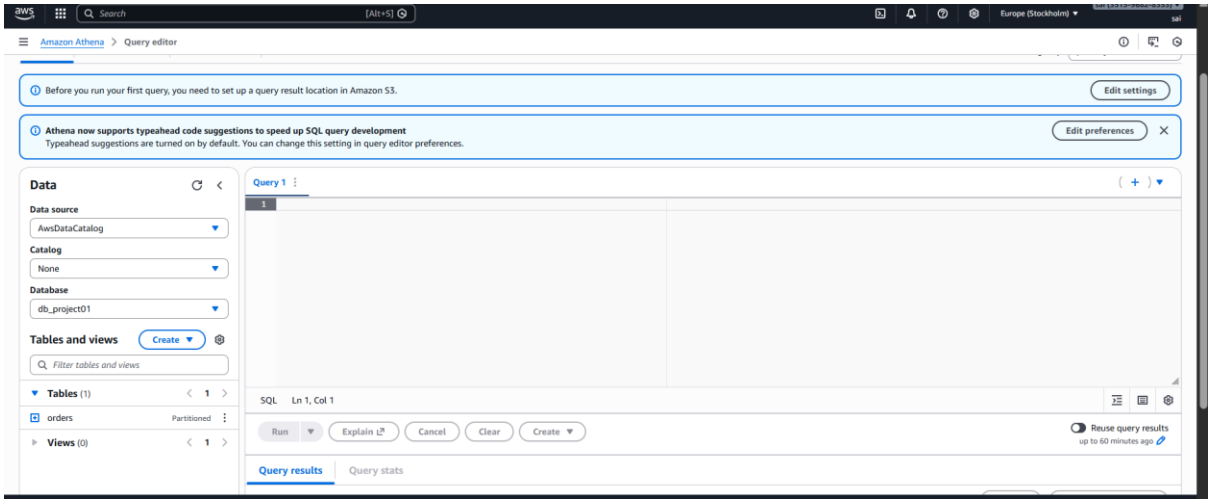
So its better to follow the format but u can follow anything as per wish

Now these partition will help us in running the queries faster as below

Conclusion: the data catelog has all the information of all the folders in the s3 location (orders)

Lets go to athena, click on query editor





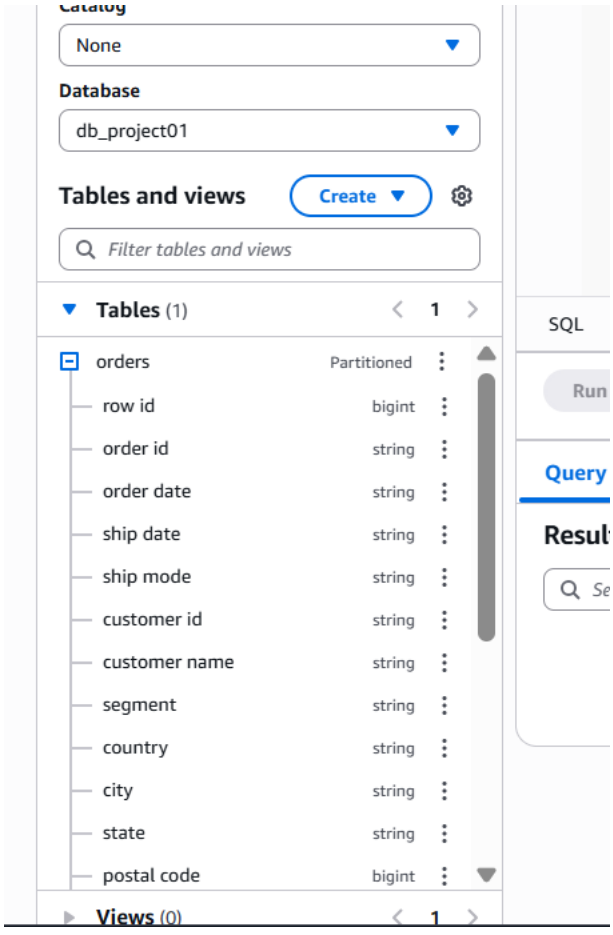
Data source:

U see AwsDataCatalog means datacatalog that I created using glue

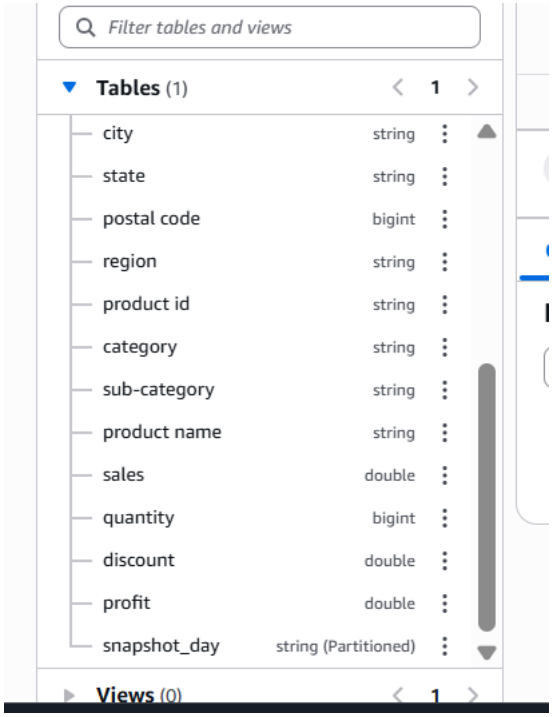
U know ur database which u created

It is also showing orders table that we created, it shows it is a partitioned table

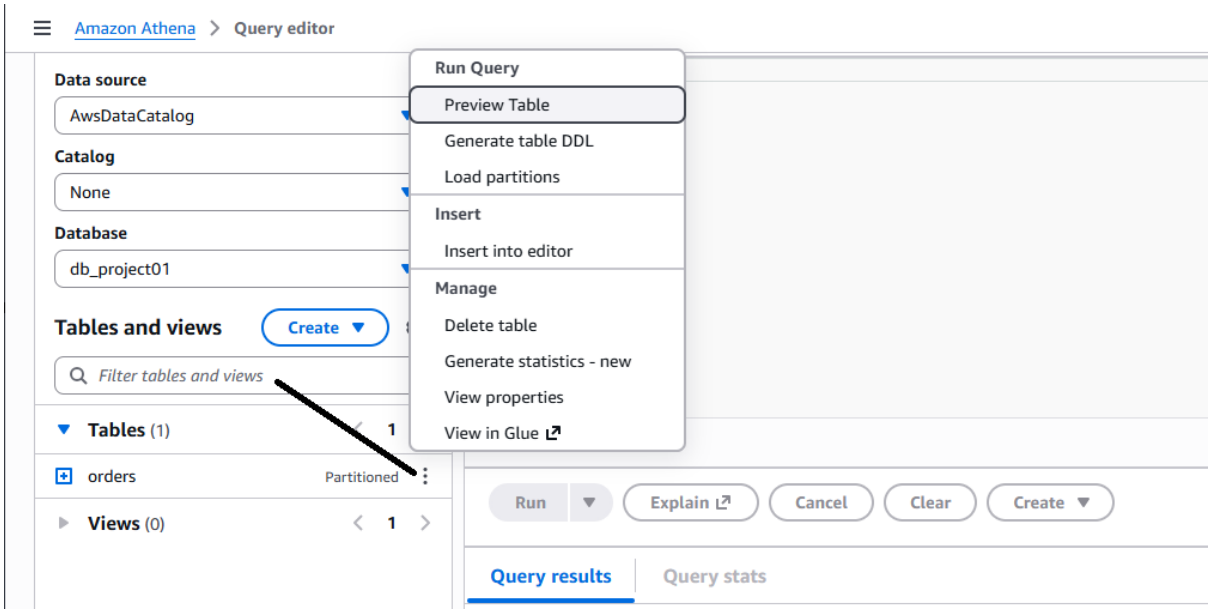
It s all the data types

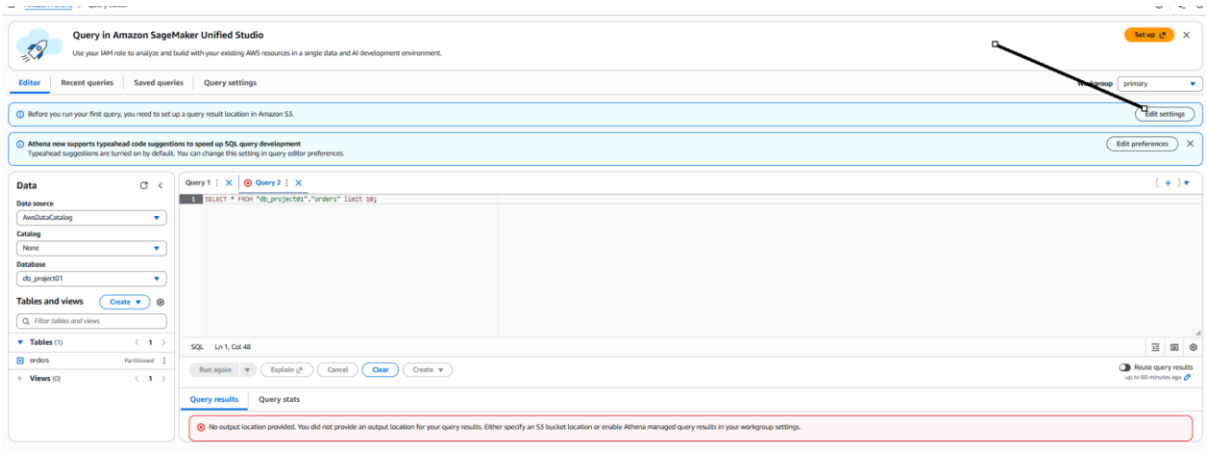


It has snapshot day column as well



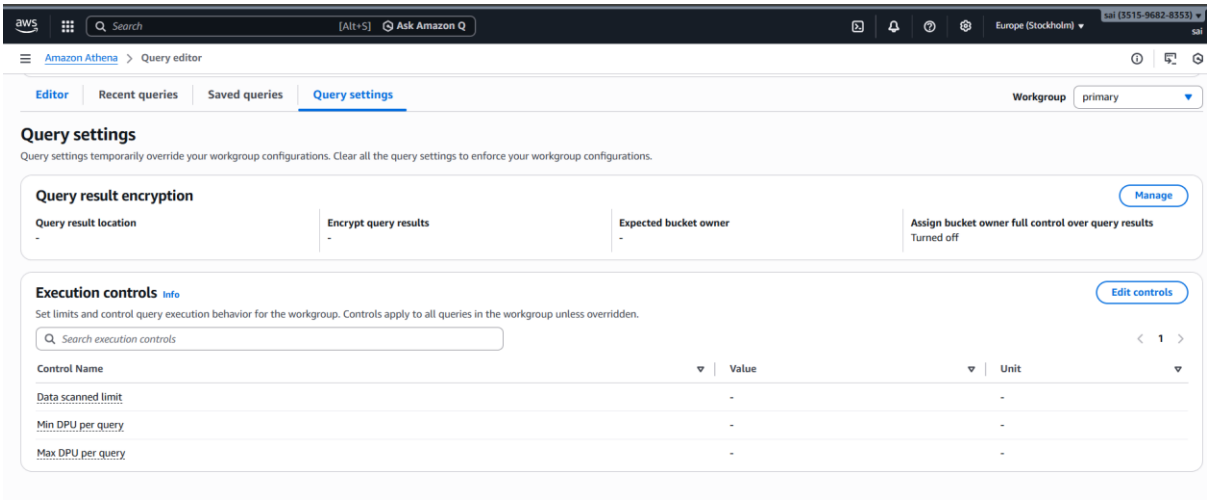
Click on preview table



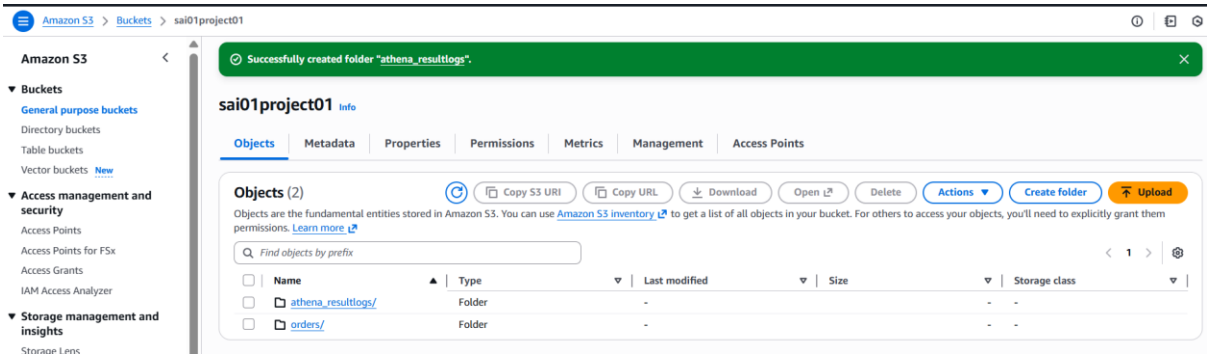


When athena runs, it stores results and that location u need to provide by above

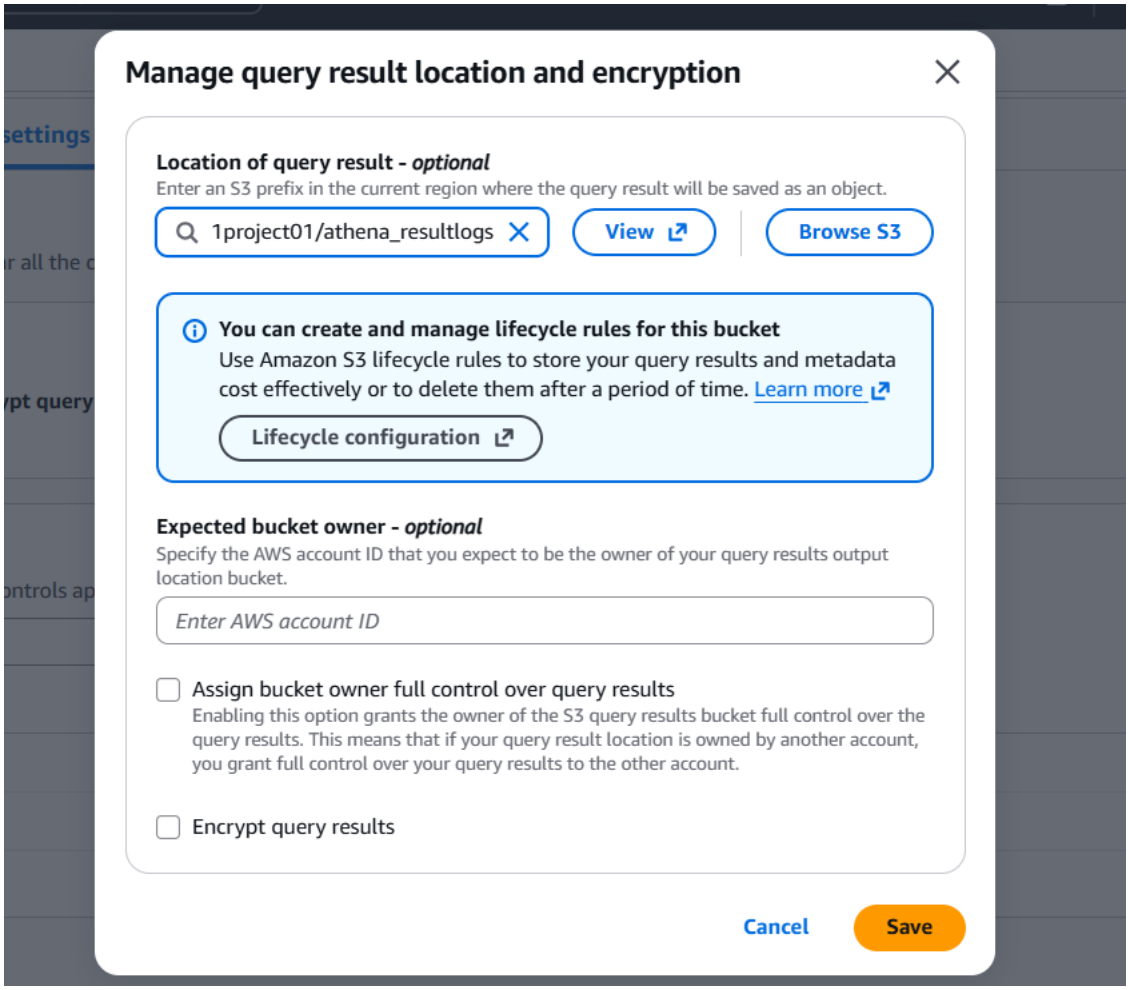
Click on query settings beside saved queries, click on manage



Before that we created a folder as shown

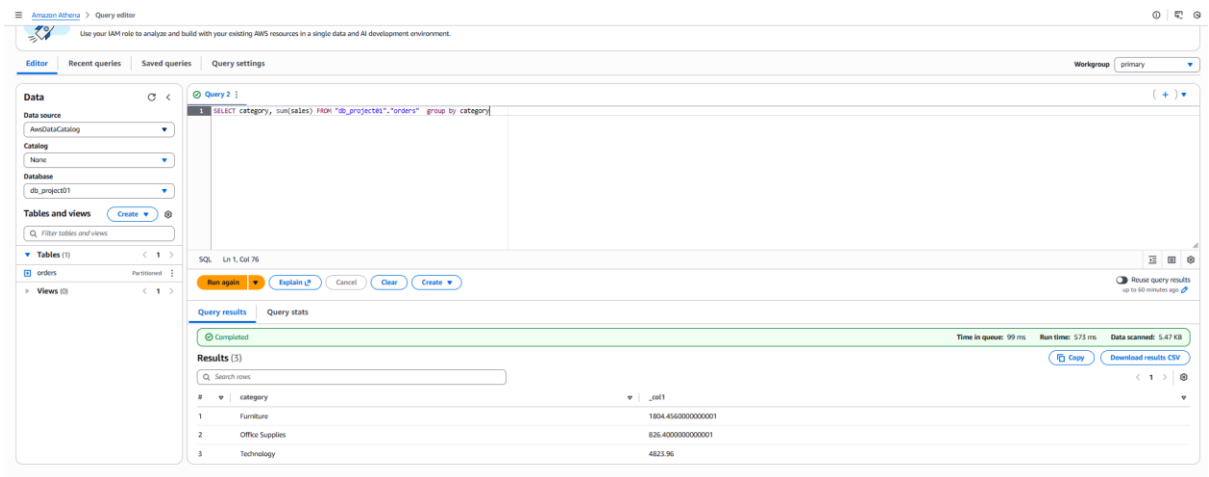


Now specify this location



Save

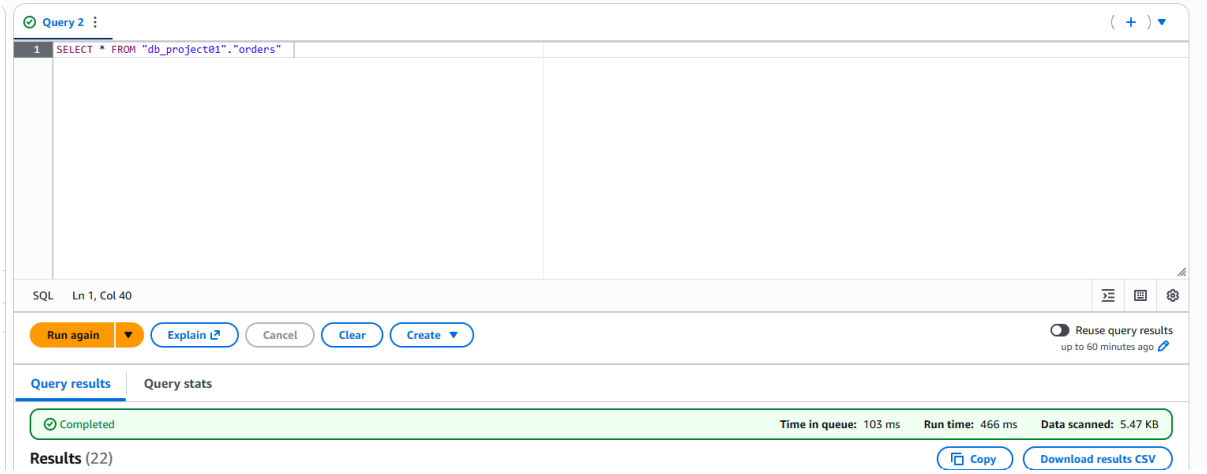
Lets go back to editor and run query



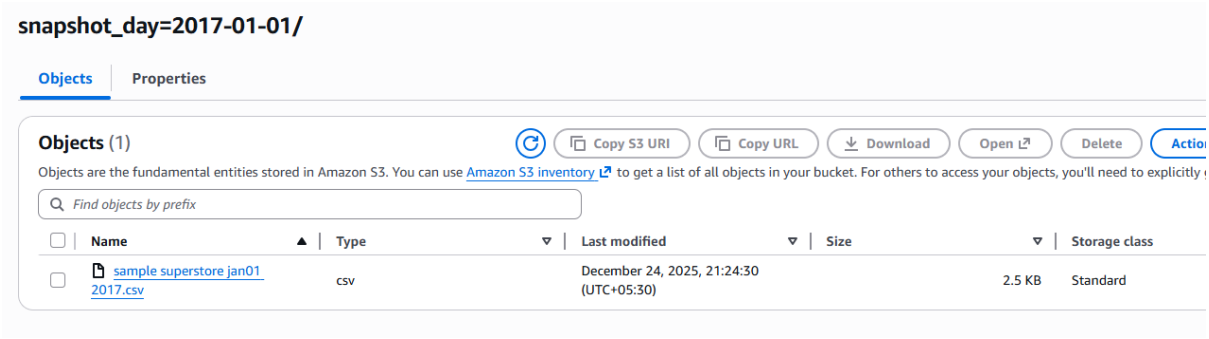
Lets see how how partitions help us

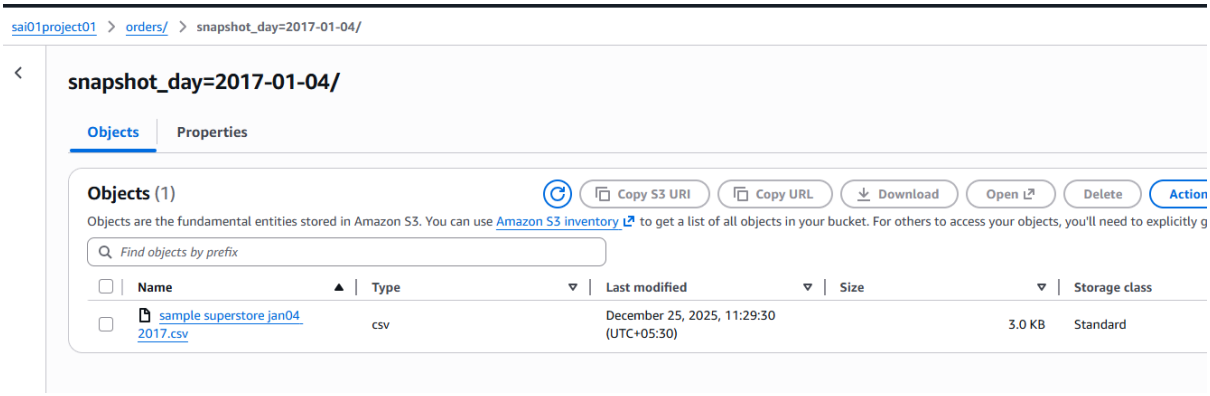
If u see above data scanned is 5.47kb

Lets run whole query



U see the data scanned is 5.47kb, it is scanning all the folders

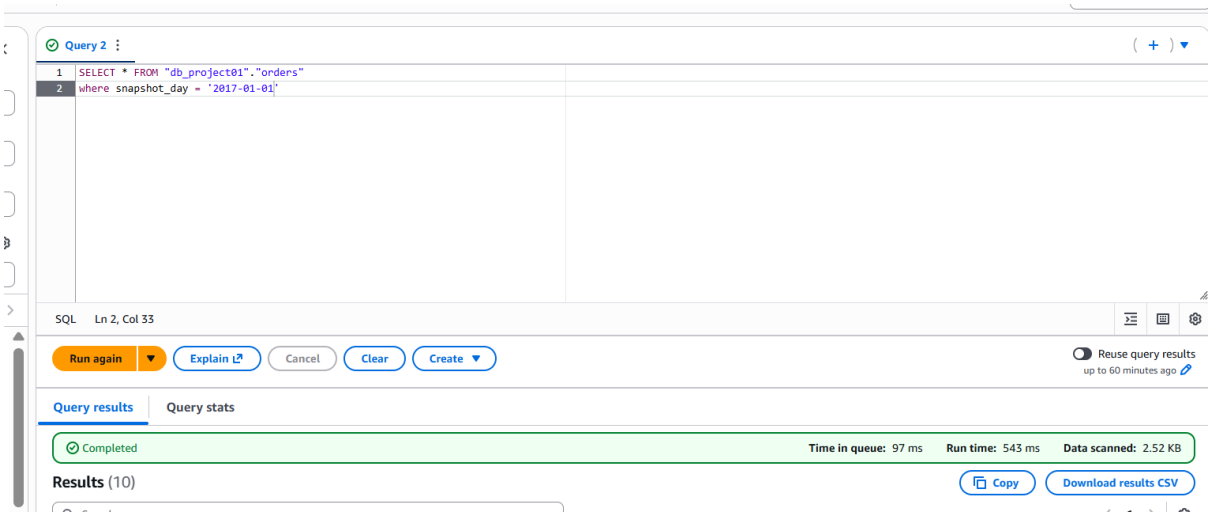




So total 2.5+3 = 5.5kb

When u query with filter as shown, u see it scanned only particular folder which is 2.52kb as per filter

In this way u can reduce scan so that it reduces ur cost and time



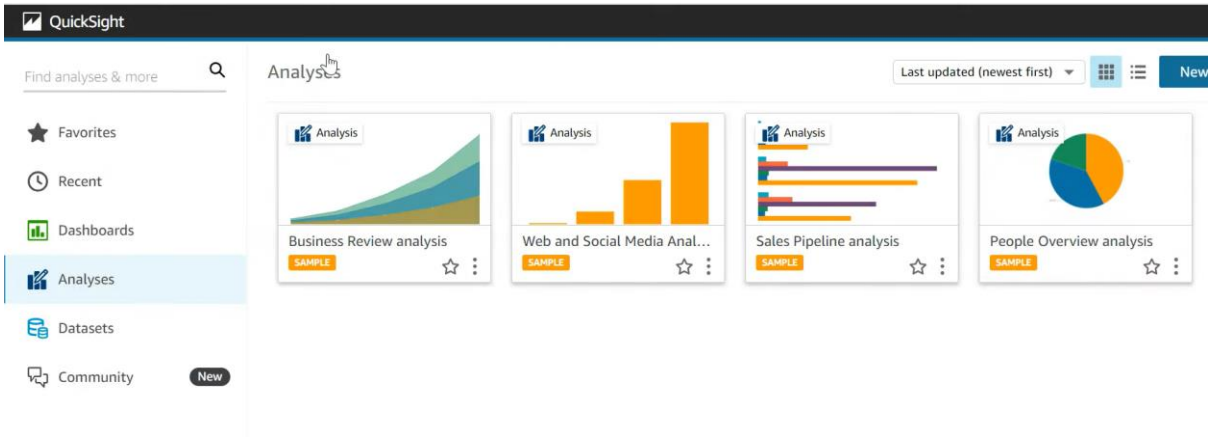
So once u created partitions, it knows where to go exactly

So try to create snapshot folders in s3 like what type of analysis u r going to do

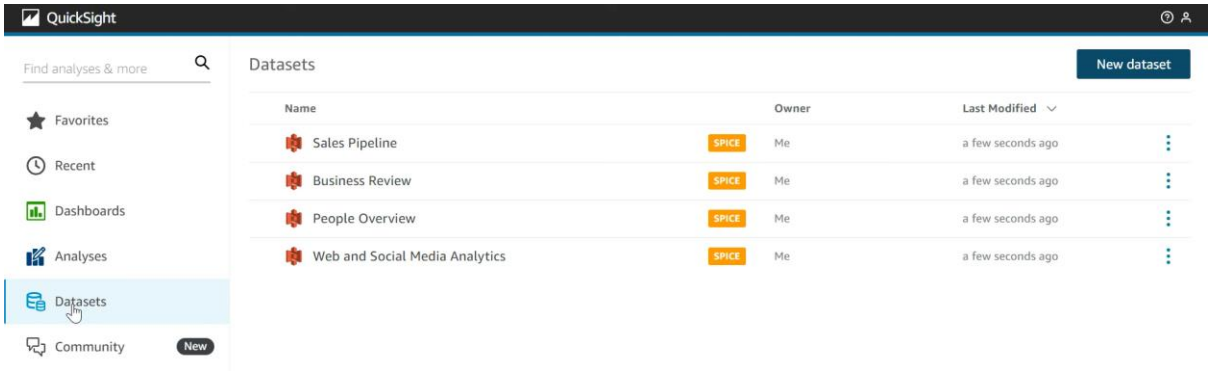
And hence while querying u can save time and cost

Lets go to quicksight

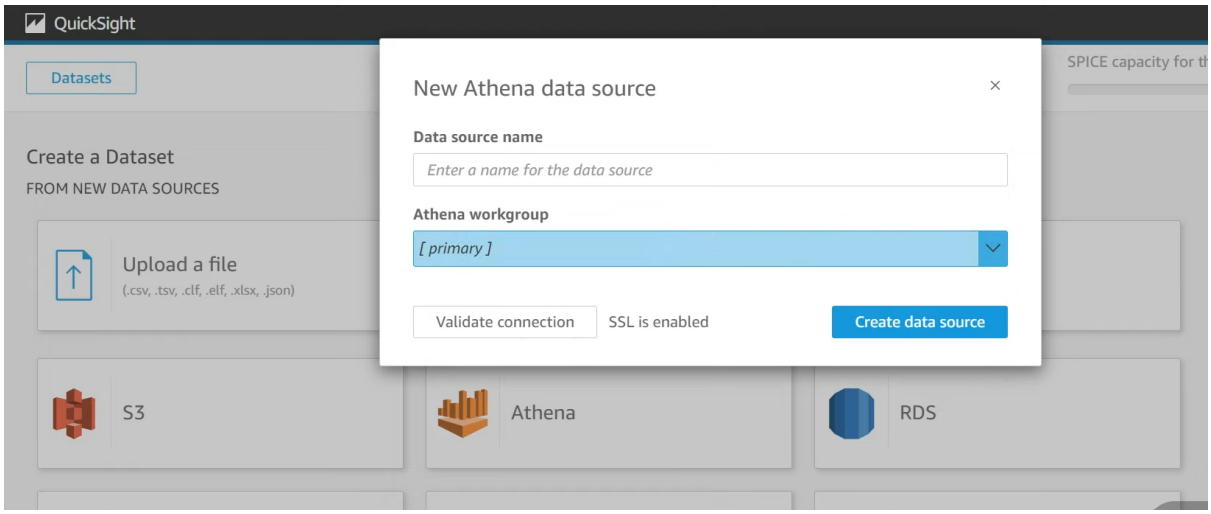
1 month free access by logging with ur email id



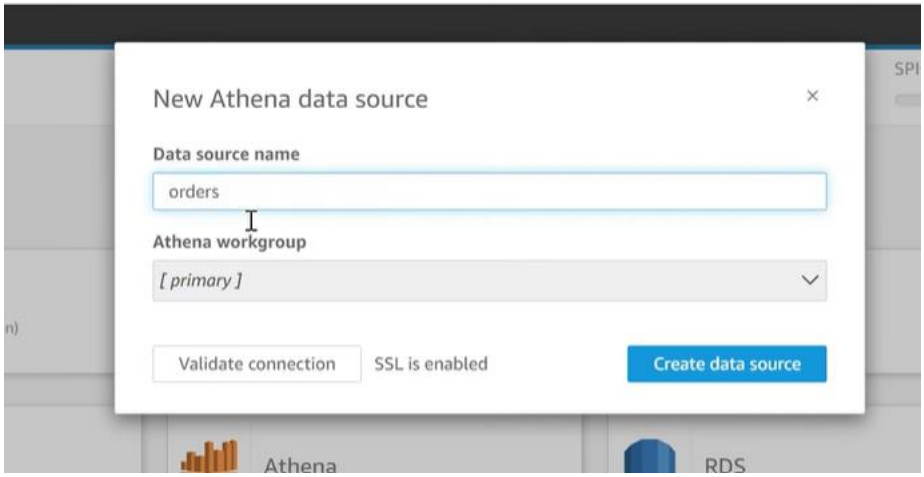
Click on dataset, click new dataset



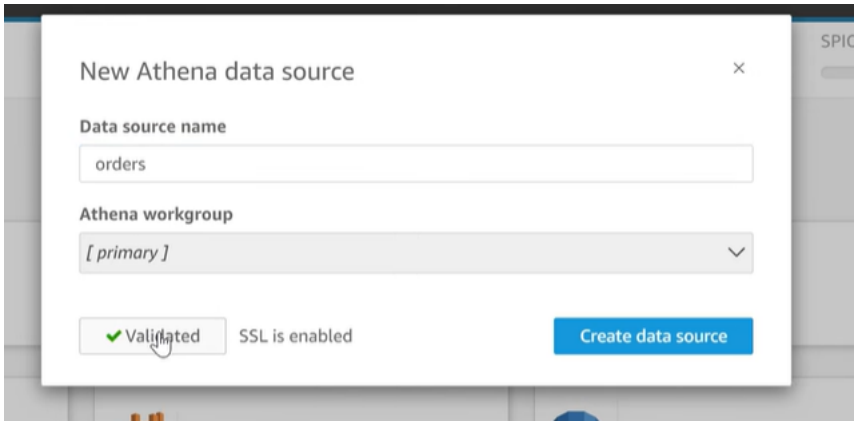
Click on athena



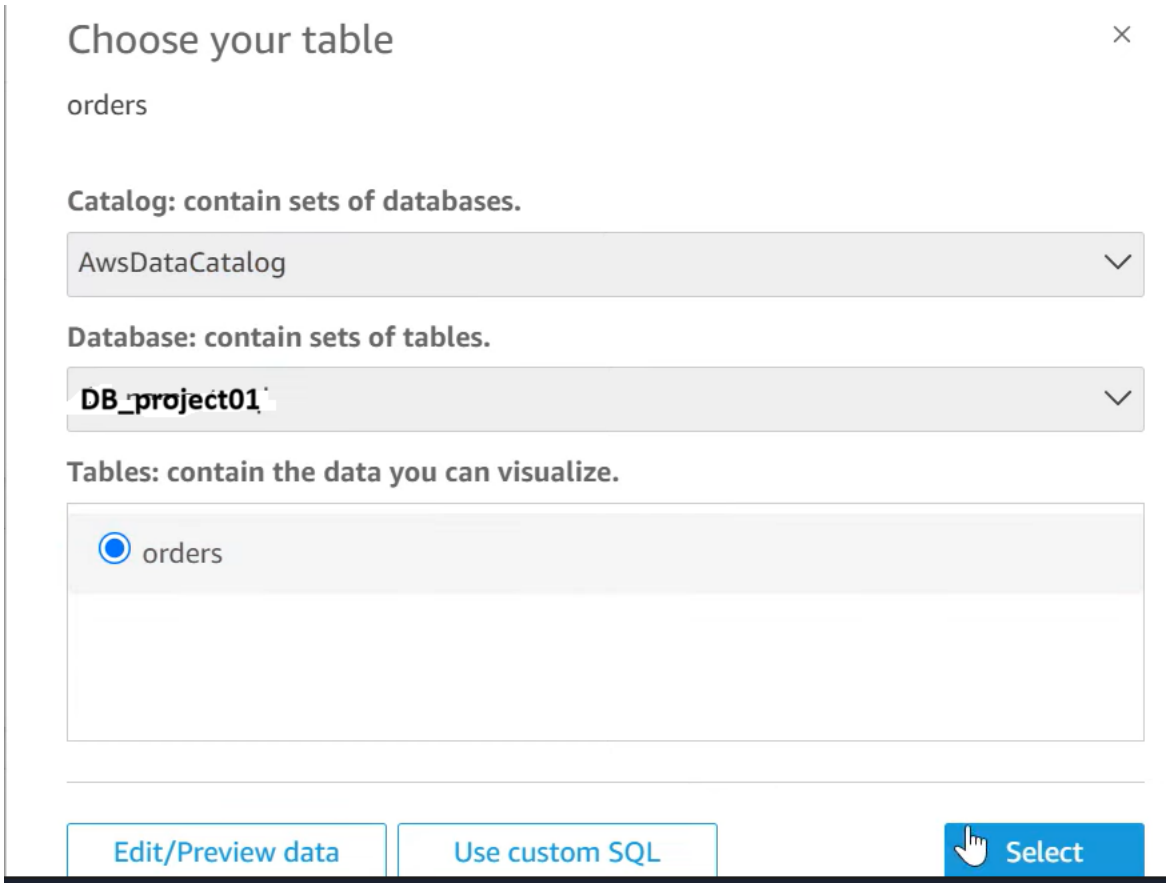
How to Install Apac... Built-in Exceptions... property tax

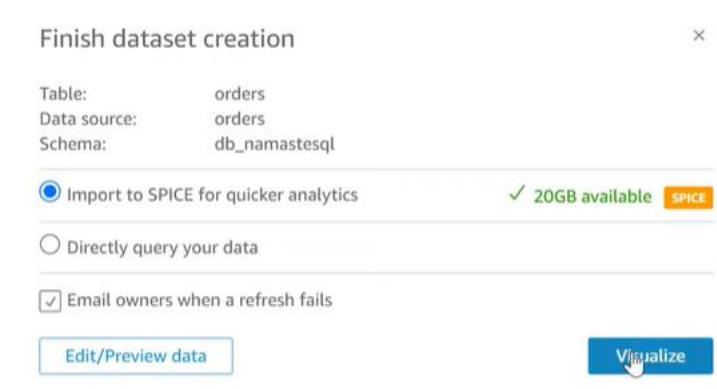


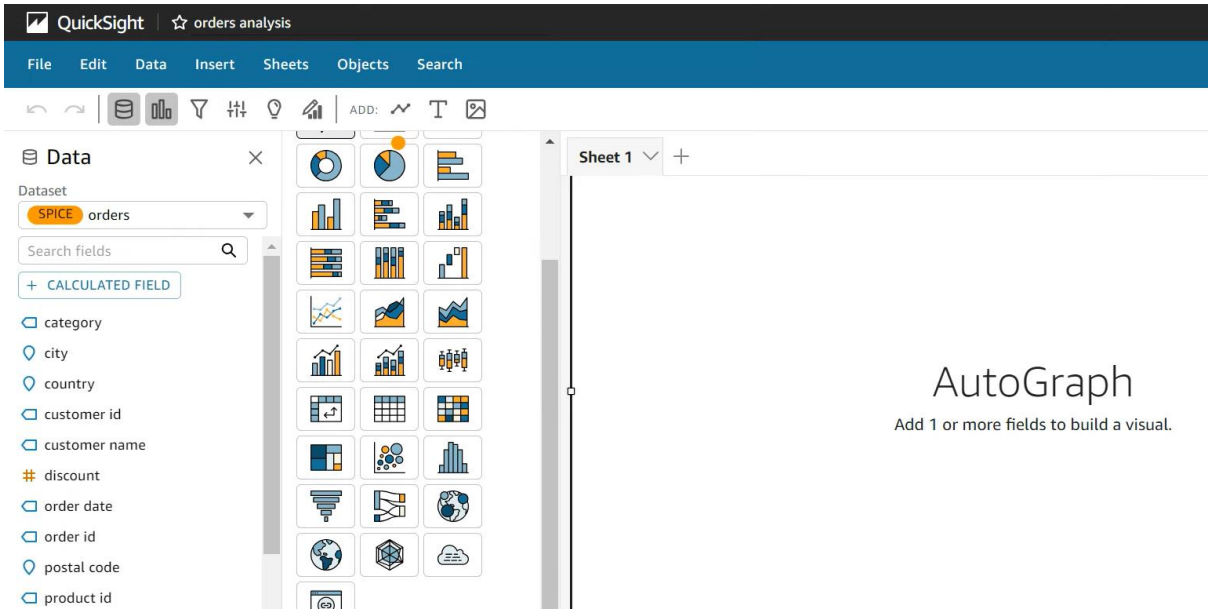
Click validate connection

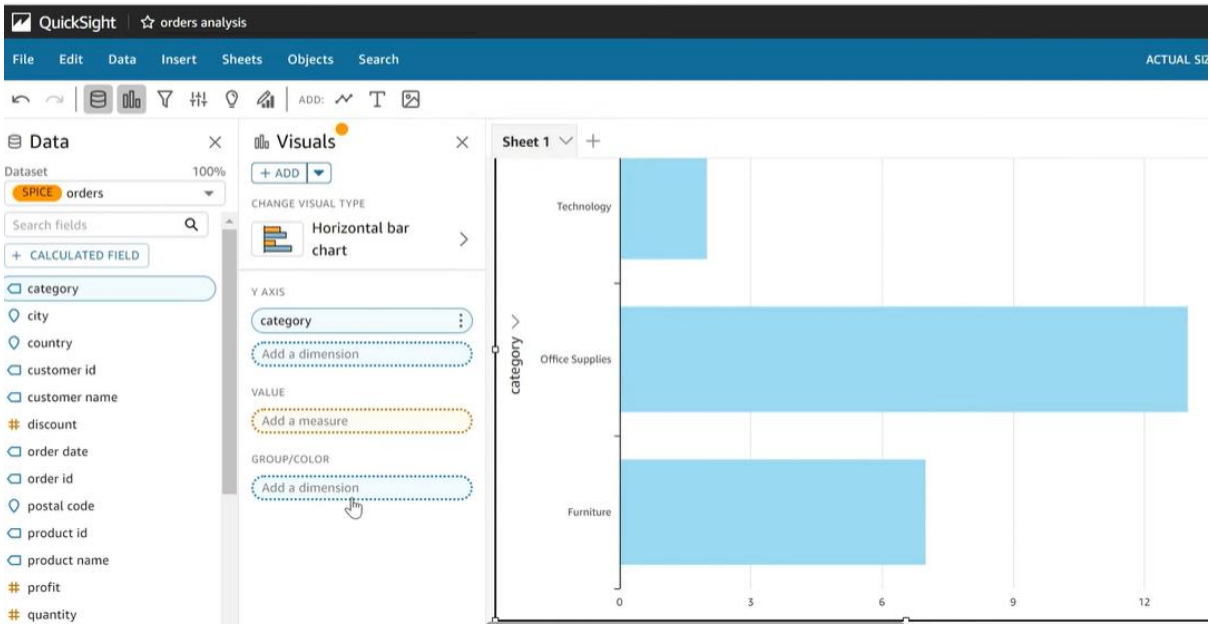


Click create data source









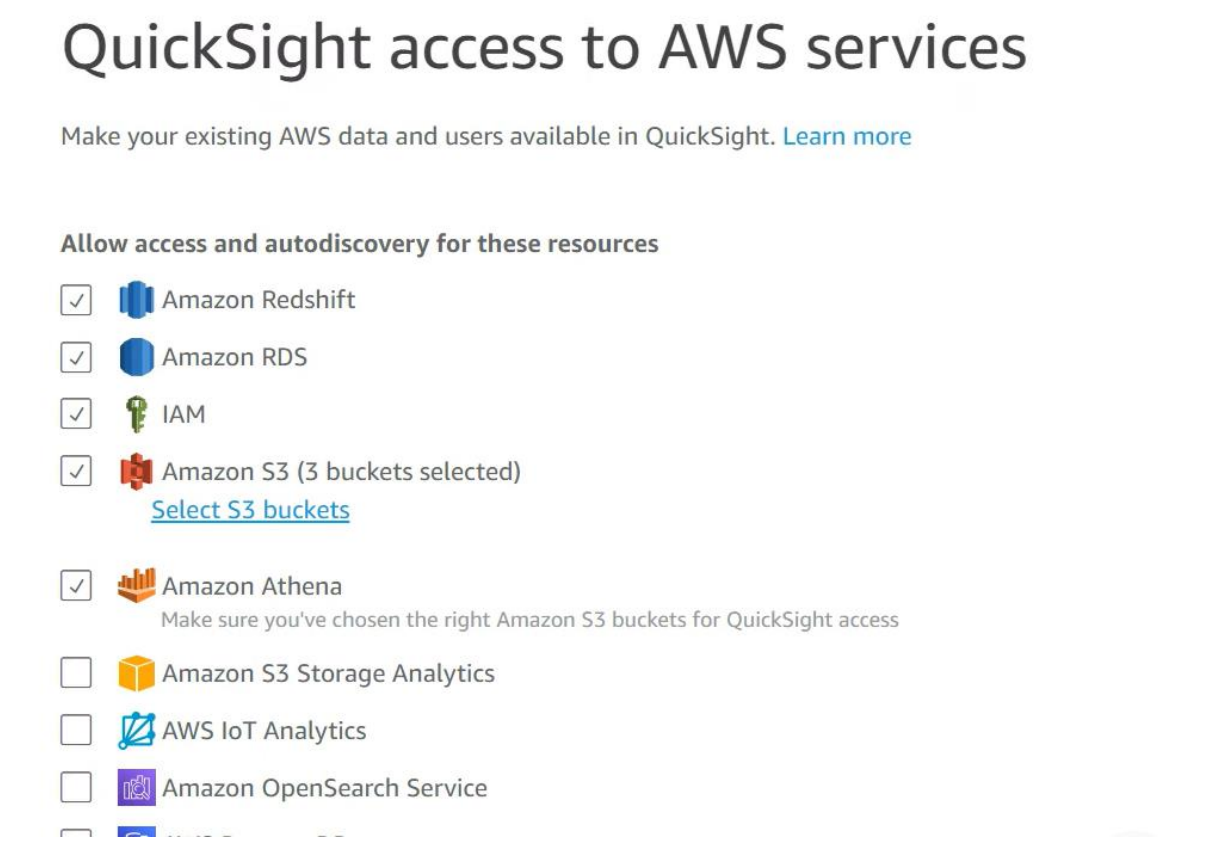
Click on horizontal bar chart,

Y axis: drag category on to y axis

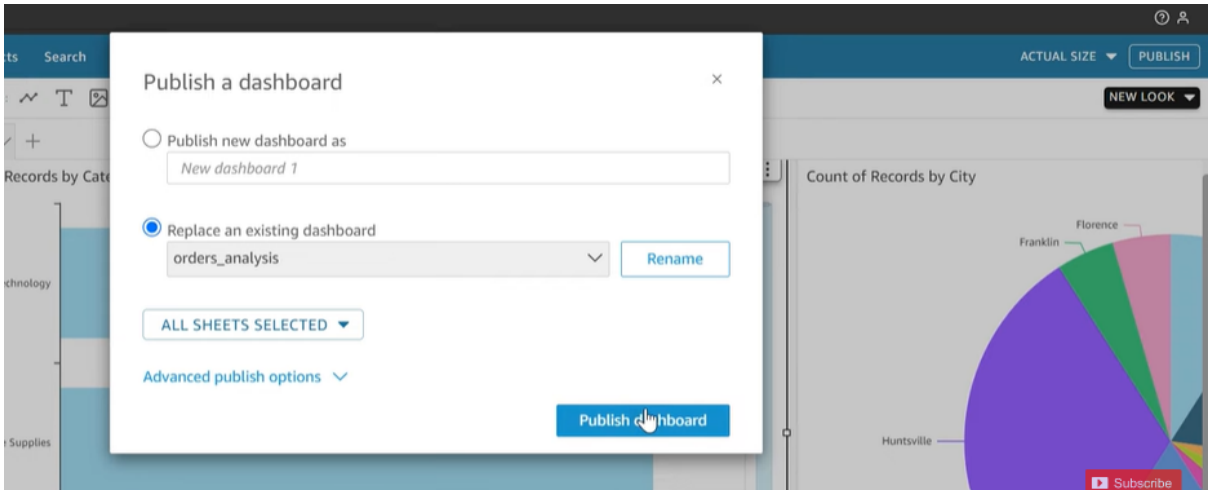
Sales into value

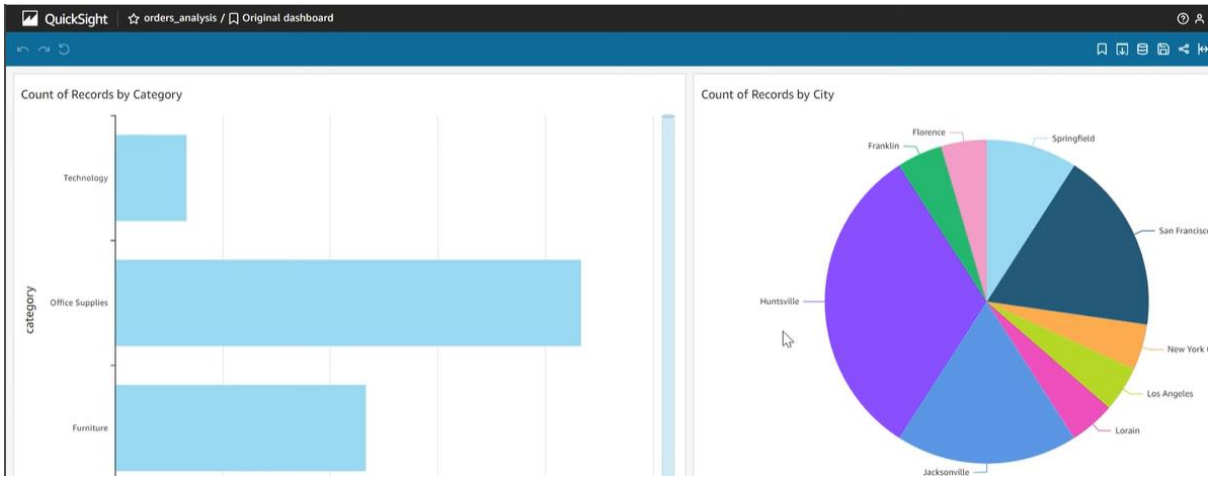
You see the graph as above

If u r unable to do the above, give access as shown



Now publish





Finally, I don’t need to store data in some db and for quick analysis I uploaded files in s3, run a crawler, run the queries on it in athena and create charts