**Build end-to-end ELT pipeline using Amazon Redshift, AWS Glue, AWS Step Function & visualize using Quicksight**

**Participant’s Guide**

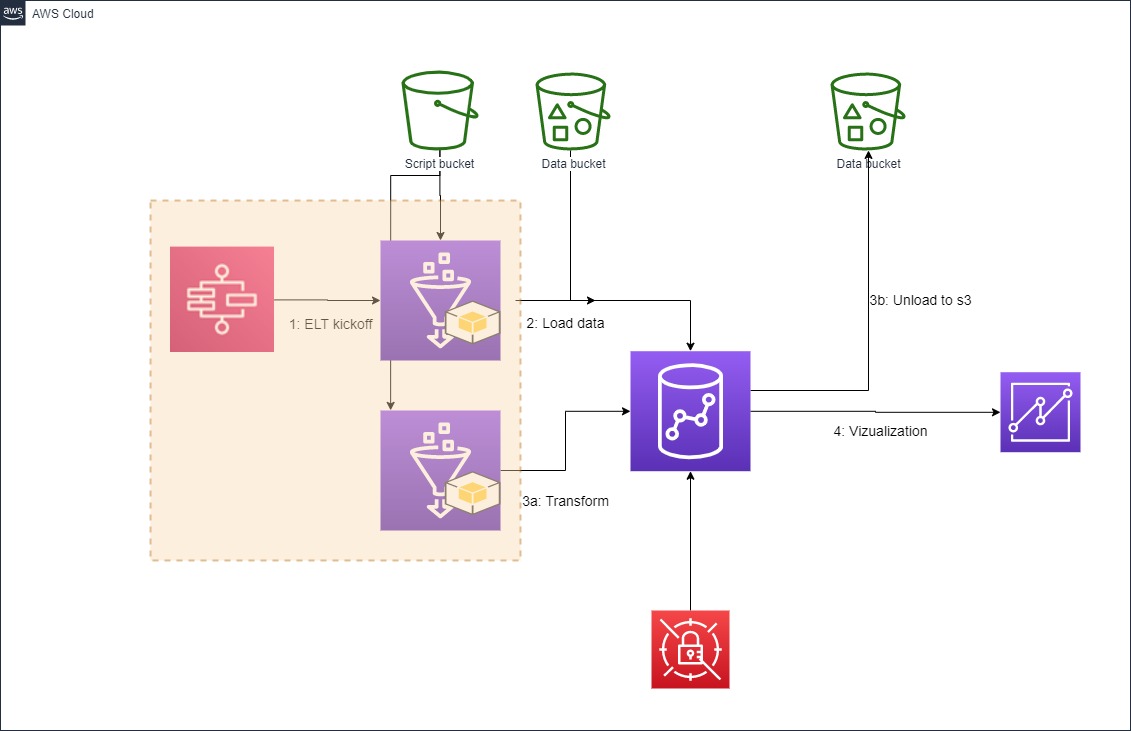
Version 1.1

*Authored by*

*Praveen Cheruvu, Director Analytics,* praveen.cheruvu@appsassociates.com

# In this lab you will create an end-to-end ELT pipeline using AWS resources to ingest and visualize [Supply Chain](https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page) sample data. This lab guides you in the process of creating AWS resources through Cloudformation stack. As you walk through the lab you will gradually get introduced to Amazon Redshift, AWS Glue, AWS Step Function, QuickSight as the building blocks for ELT pipeline.

**Lab –Architecture Diagram:**



**Please use us-east-1, us-east-2 and us-west-2. Instructions may not work in other regions.**

**Cloud Formation**

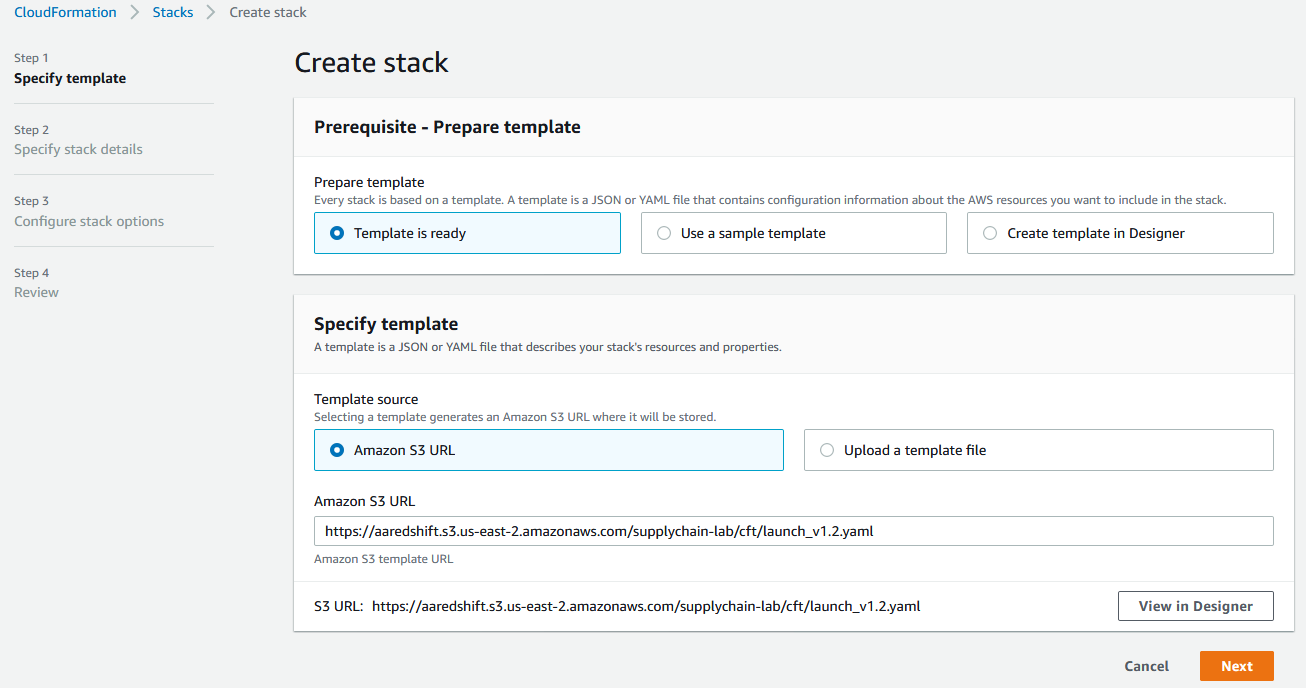
To launch this Redshift cluster, Glue resources, Step function we use Cloud formation template.

After you click on the Launch Stack button below follow the screenshot instructions to avoid any S3 access errors.

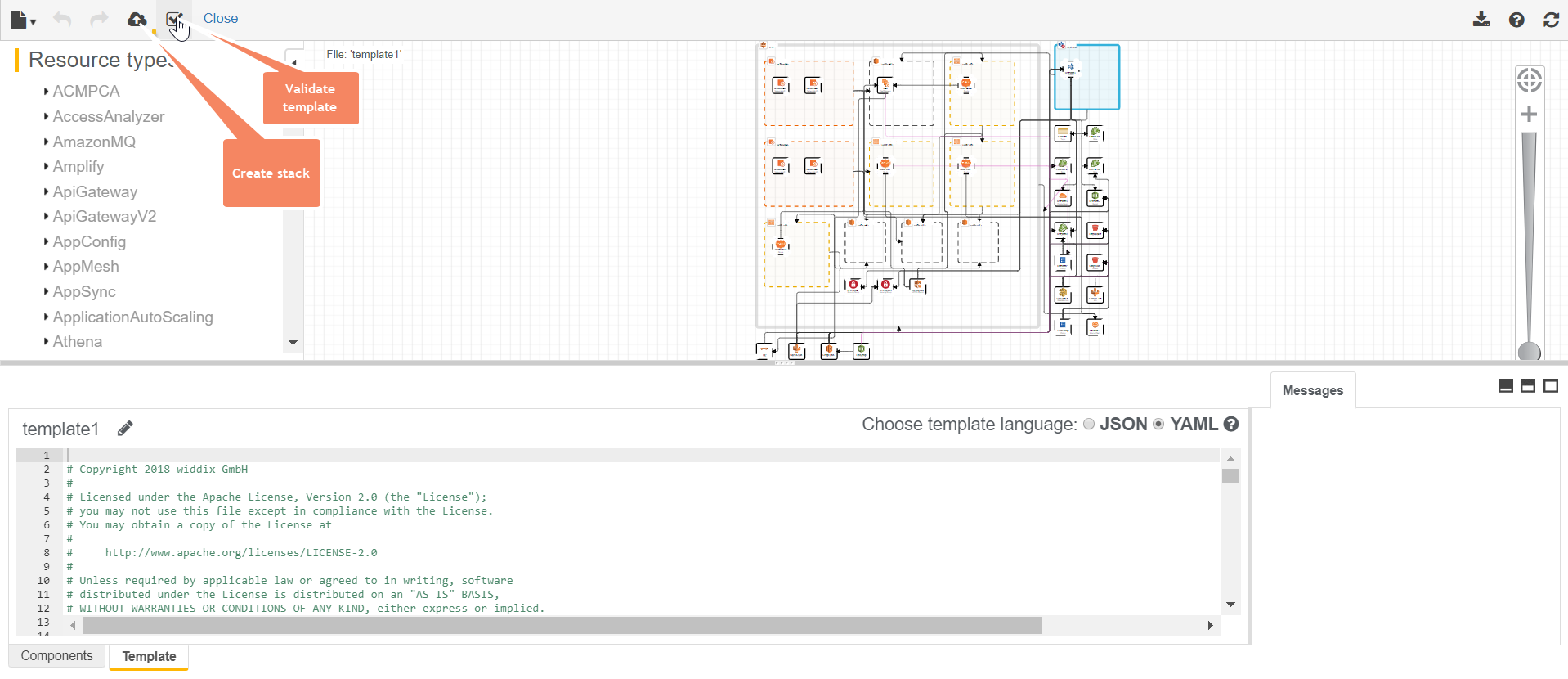
* Click the below button to launch necessary AWS resources required for this lab. Please note you need to have an AWS account prior to launch these resource.

[](https://console.aws.amazon.com/cloudformation/home?#/stacks/new?stackName=SupplyChain&templateURL=https://aaredshift.s3.us-east-2.amazonaws.com/supplychain-lab/cft/supplychainlaunch_v1.2.yaml)

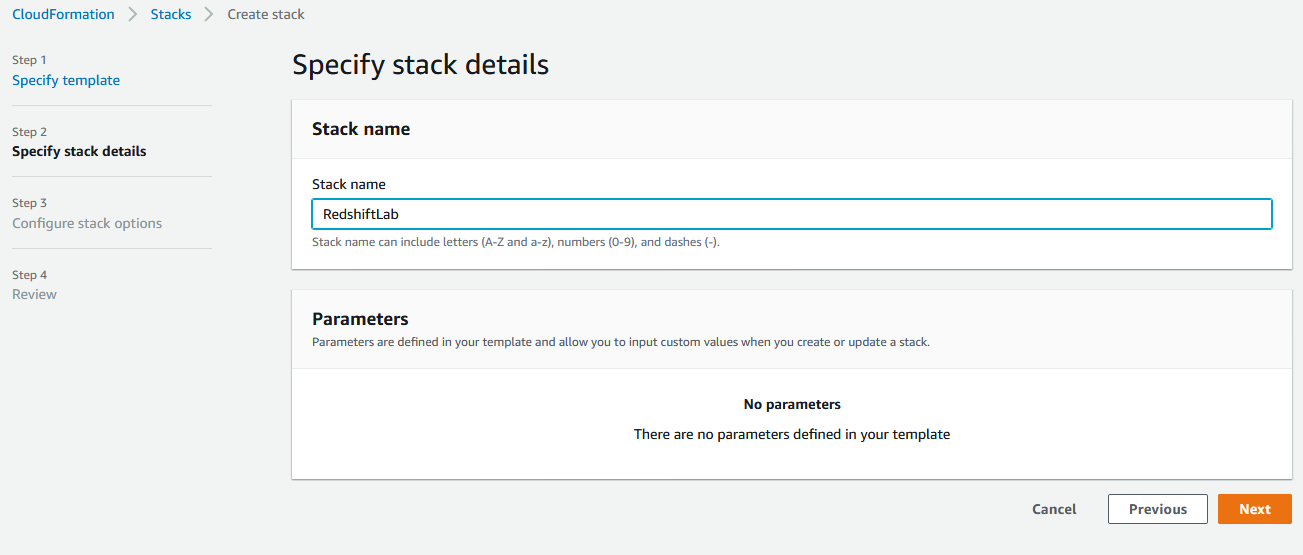
* Click **View in Designer** button.



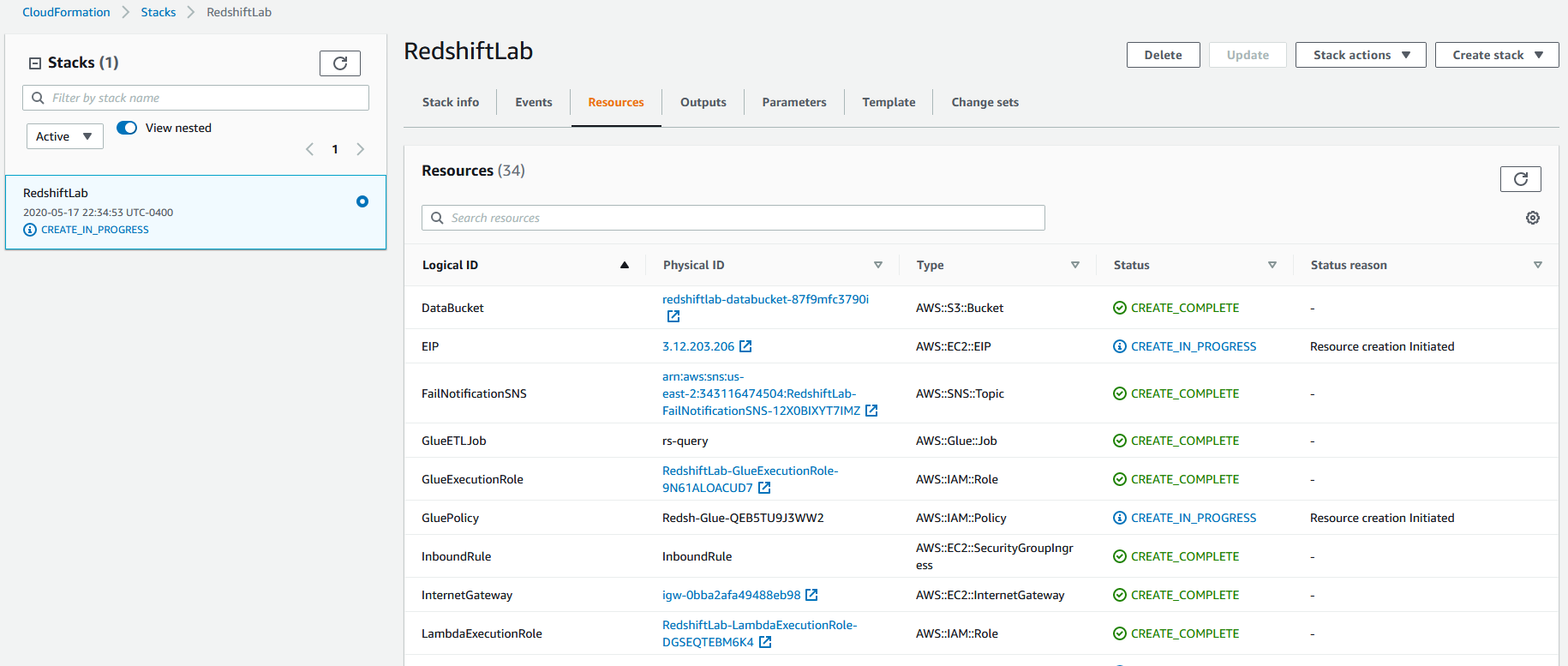
* Wait for 5 seconds to allow the architecture diagram populated. Once the diagram becomes visible click on **Validate template** button. One validated click on **Create stack** button.



* Click Next on **Create stack** page.
* Give a name to your stack like “RedshiftLab” and hit Next.



* Hit Next one more time after which you are in Review page. Scroll down to bottom, select **I acknowledge …** and hit **Create stack**.
* For the next 5 minutes monitor progress of your Cloudformation stack.



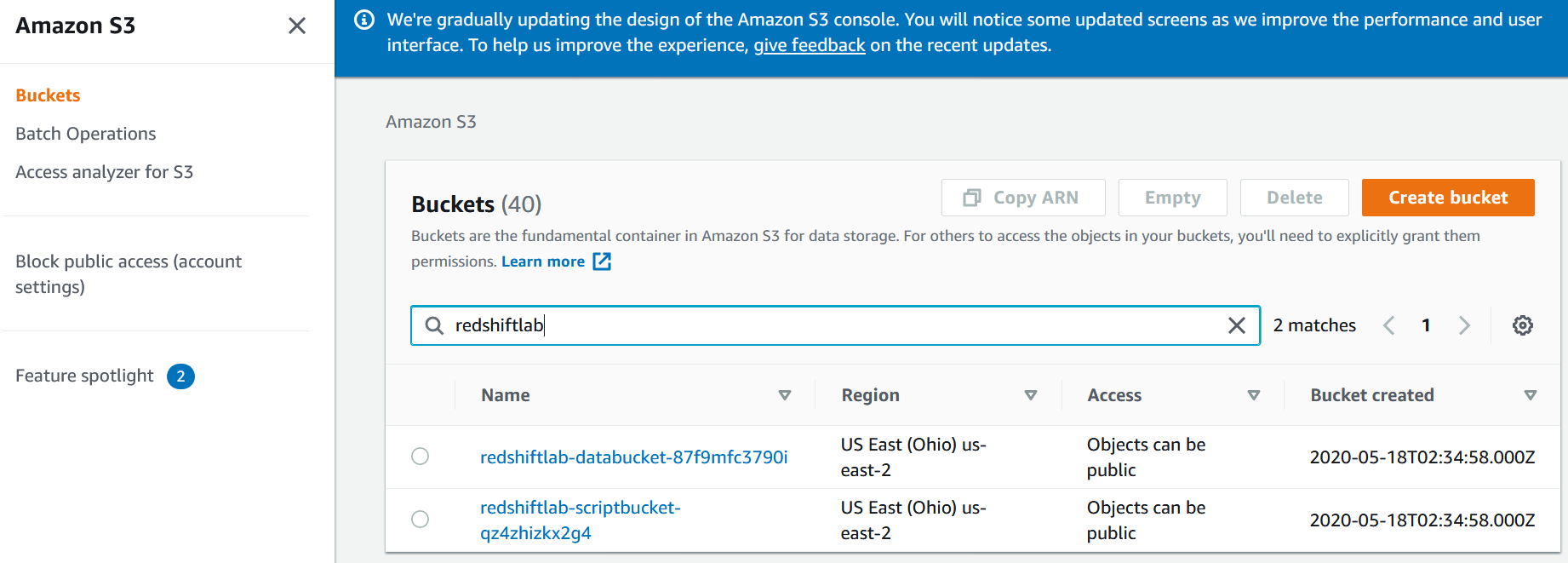
* After about 5 minutes the Stack creation completes.

# **Setting up:**

In this lab a Python Shell job is the heart of all the ELT being done in Glue. This Python scripts and other SQL scripts are required to be placed in your lab specific s3 bucket.

The very first thing you will do after logging in into your team AWS account is to navigate to S3 service.

Navigation: **Services > S3**



You will see 2 buckets already created for you- one is <stackname>-scriptbucket and another is <stackname>-databucket.

Next you will replace files in each of these buckets. Follow below steps.

Note: If you are not familiar working with s3 console please ask the instructor how to perform create folder, upload files etc.

* First download necessary files from below locations into your local machine.
  + python <https://github.com/praveen05ch/redshift-lab/tree/SupplyChain/python>
  + sql <https://github.com/praveen05ch/redshift-lab/tree/SupplyChain/sql>
  + data <https://github.com/praveen05ch/redshift-lab/tree/SupplyChain/data>
* Upload these downloaded files into your lab specific s3 bucket in proper location as shown below-
  + [python](https://github.com/saunakc/redshift-dataengineering-lab/tree/master/python) 🡪 redshiftlab**-scriptbucket**-{randomstring}/python
  + [sql](https://github.com/saunakc/redshift-dataengineering-lab/tree/master/sql) 🡪 redshiftlab-**scriptbucket**-{randomstring}/sql
  + [data](https://github.com/saunakc/redshift-dataengineering-lab/tree/master/data) 🡪 redshiftlab-databucket-{anotherrandomstring}/csv

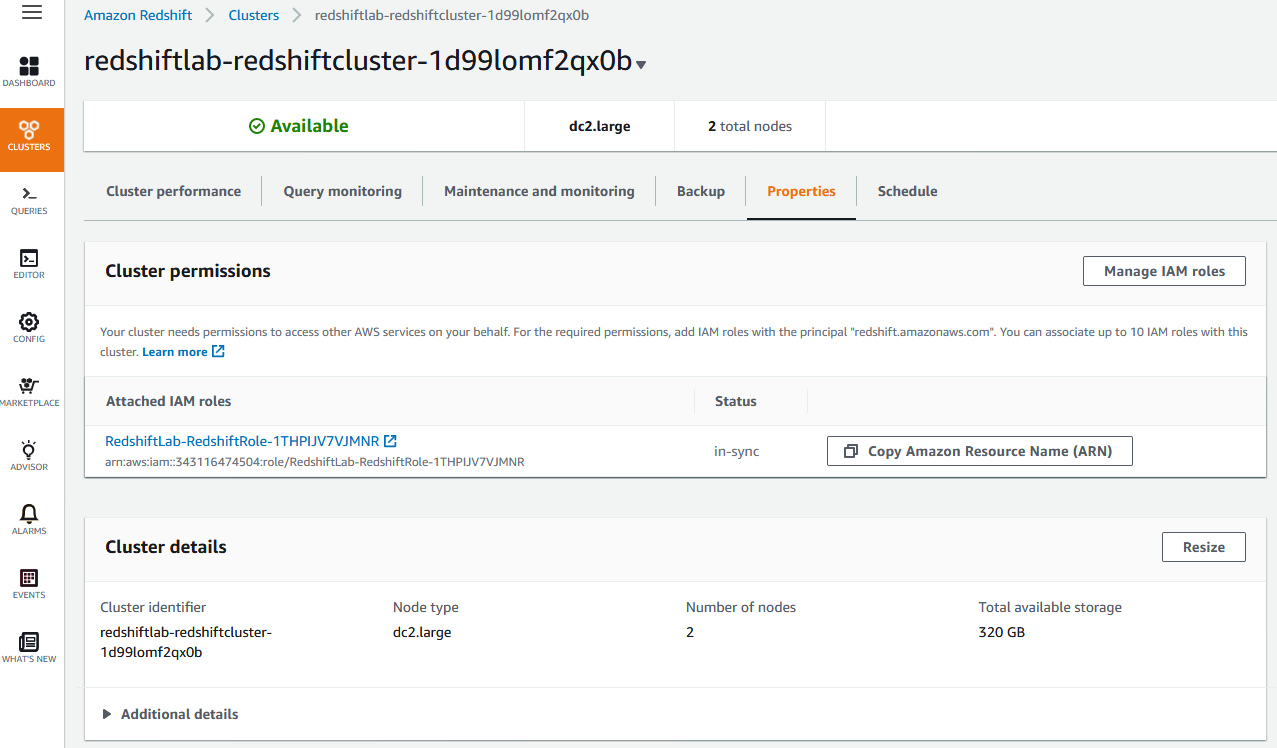
# **Navigate to your Amazon Redshift cluster:**

Navigation: **Services > Amazon Redshift > Clusters**.

For this lab a 2 node dc2.large Redshift cluster has already been crated for you.

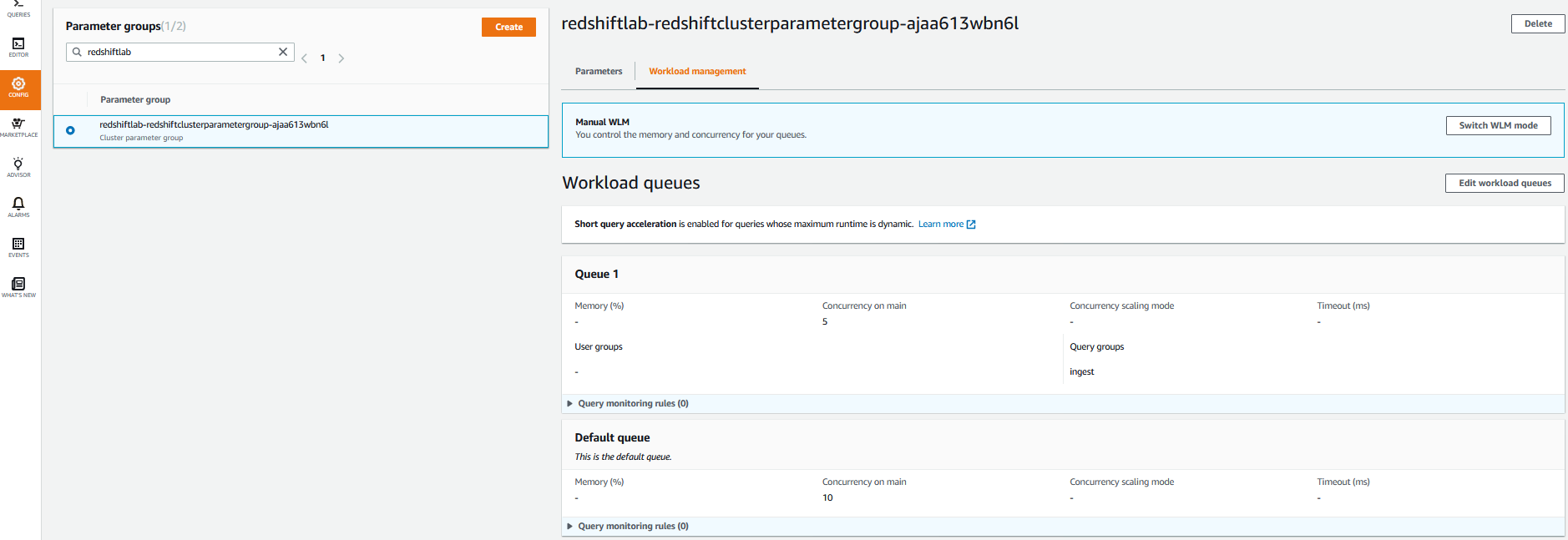
Review and familiarize yourself with the new Redshift console. Specifically check out the **Dashboard**, **Editor** and **Config** icons from the left pane.

Under Dashboard click on the single cluster



Click on the cluster and then **Properties** tab.

Under **Database configurations** click on the parameter group and review the WLM configuration.



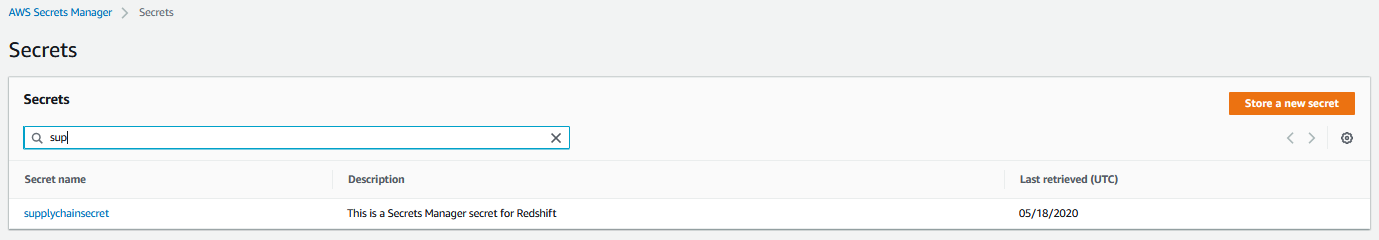
**Summary of the WLM:**

* Manual WLM with a total of 2 queues
* Queue 1 has concurrency of 5 and query group = “ingest”
* The 2nd queue is the Default queue with a concurrency of 10.

## Retrieve Redshift password

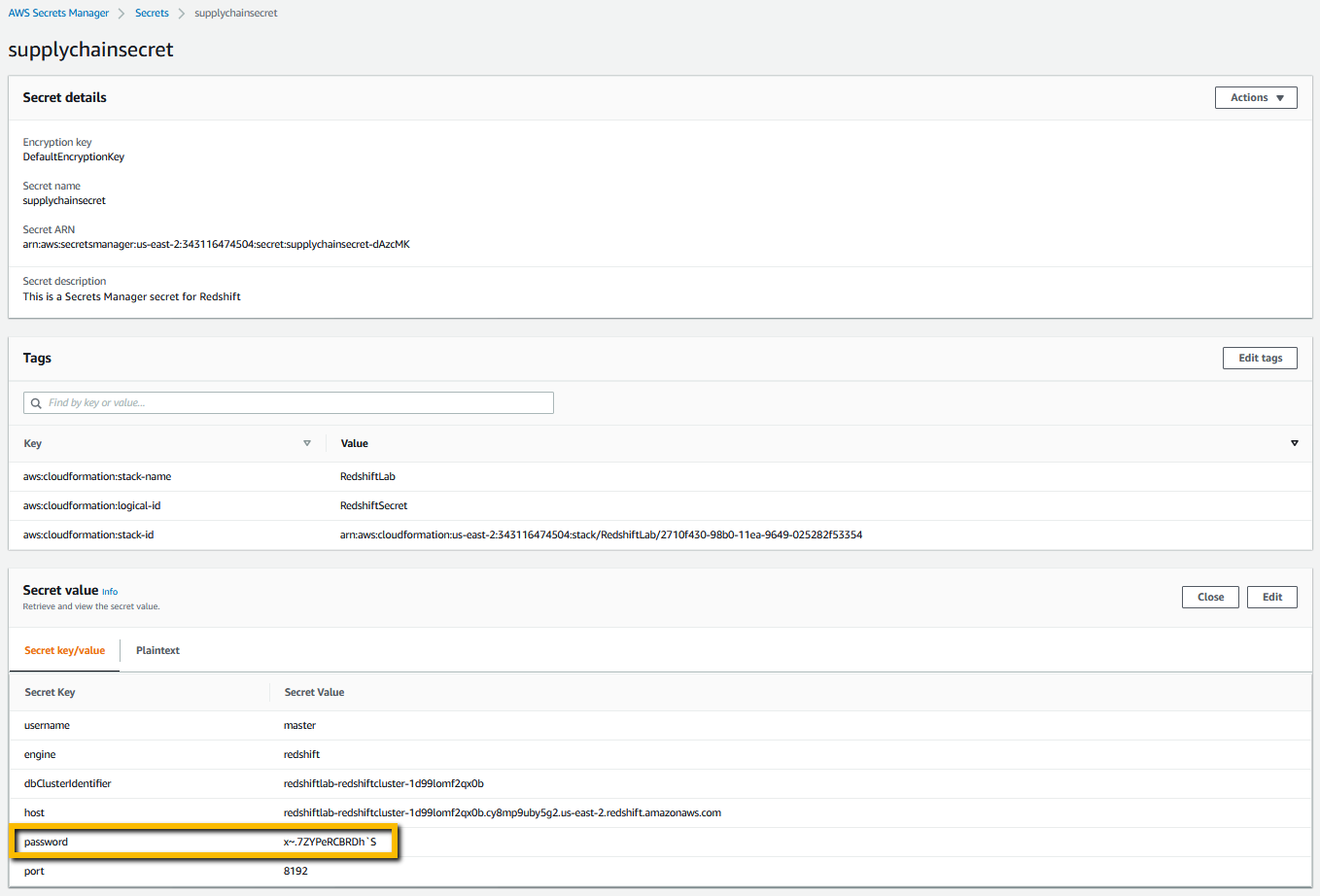
The login info for the Redshift cluster is to be stored in AWS service **Secrets Manager**.

Navigation: **Services > Secrets Manager**.



Click **supplychainsecret**

Click on **Retrieve the secret value**

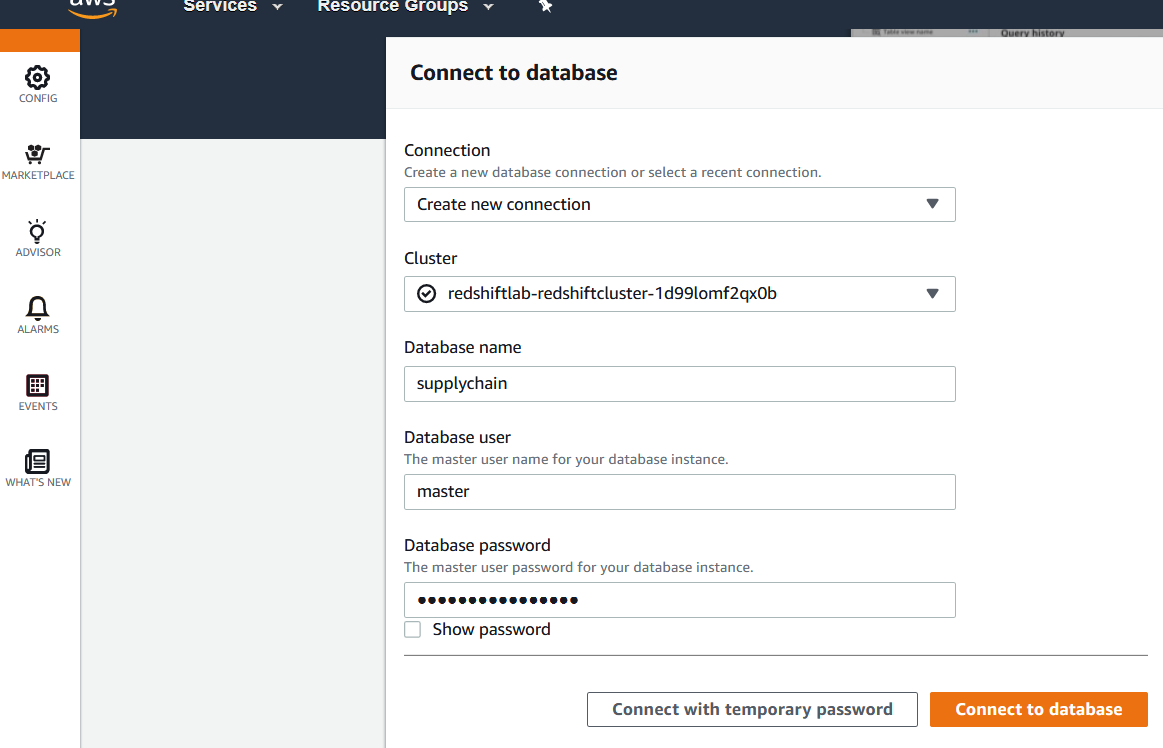


## Connect to the Redshift cluster

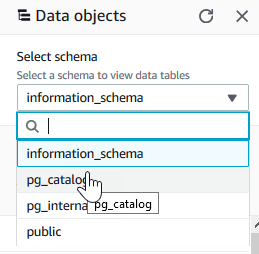
Navigate to **Services > Redshift**

Click on **Editor** from the left pane. In the **Connect to database** prompt enter the details for your cluster and connect to database.

* Database name: supplychain
* Database user: master
* Database password: (You should have retrieved from the Secrets Manager described under section Retrieve Redshift password)



Under **Select schema** click on drop down. There will be only 4 schema at this point.



# **Create ELT pipeline:**

We will create Glue job as the building block for ELT pipeline. The pipeline will be created using State machine to orchestrate between jobs in an ELT flow.

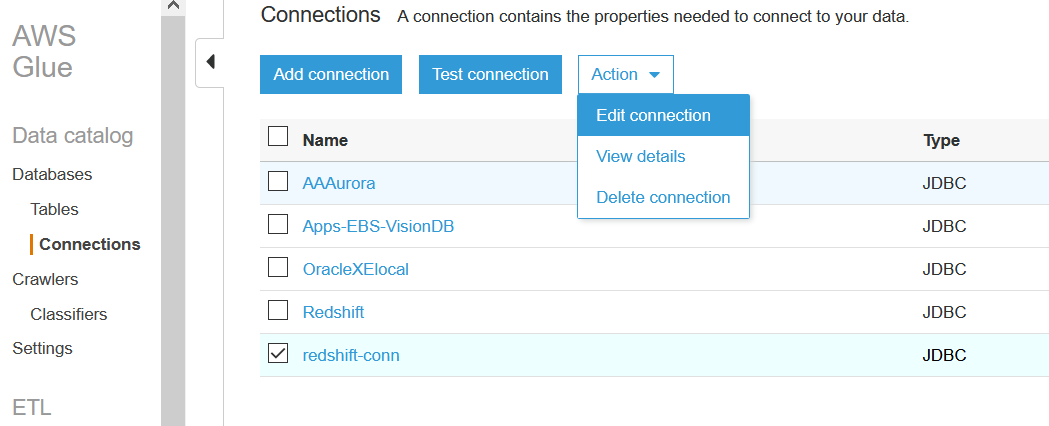
Note: NO code writing is required. The necessary code is included in the lab package. However you can use this code to build your real world pipeline.

## Glue Job

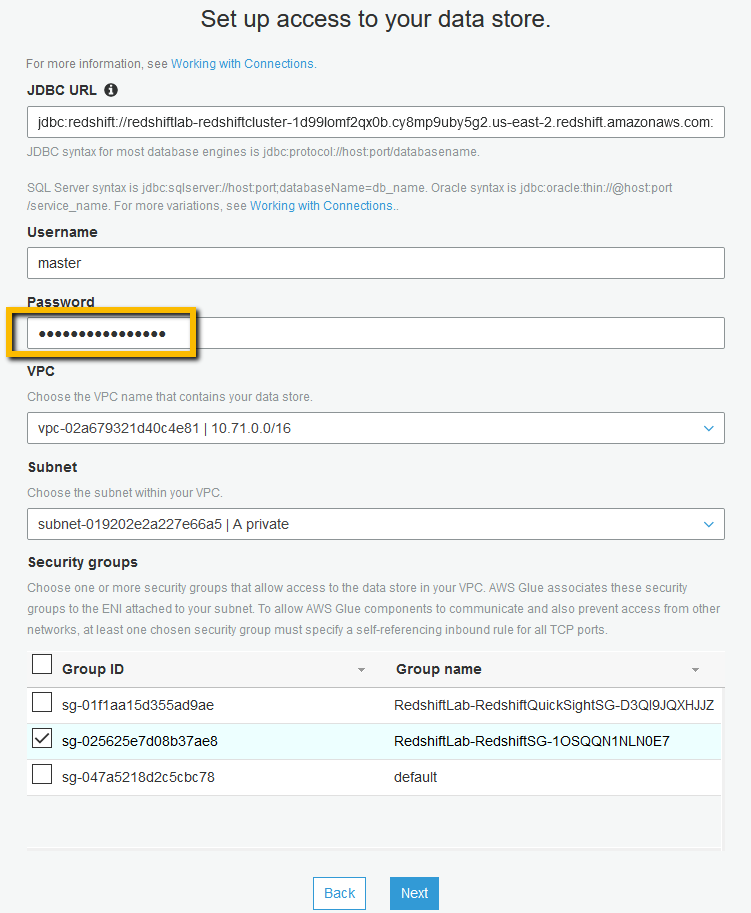
Navigate to AWS Glue service by **Services > AWS Glue**

Click on **Databases > Connections**

A connection name “redshift-conn” should exist. Select the connection and **Action > Edit connection**



Click Next until you reach the **Set up access to your data store**. window.



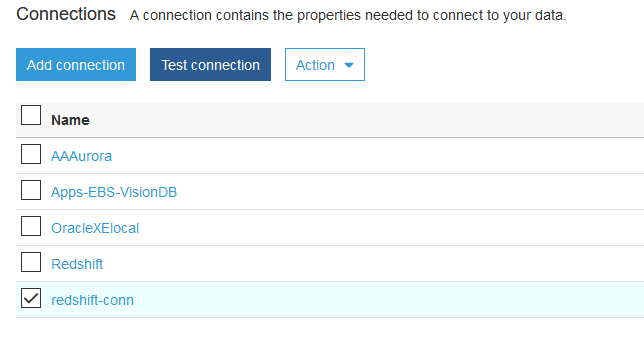
Here you need to enter the connection password. Find the Redshift password as described under section “Retrieve Redshift password”.

Click **Next** and

Click **Finish**.

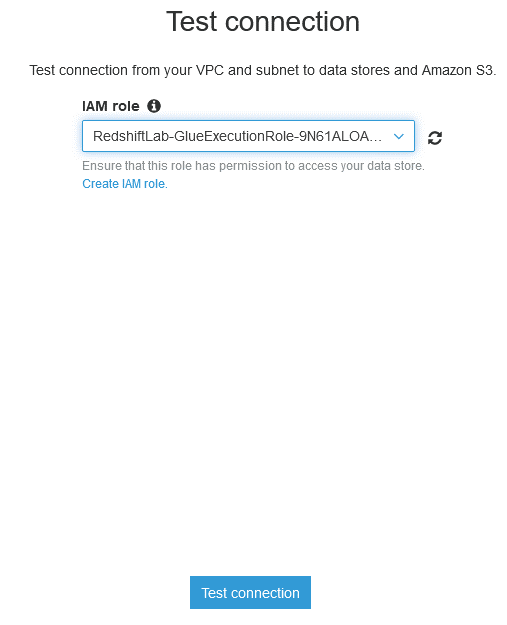
Come back to the Glue 🡪 **connection** screen

if you are not already and select the “**redshift-conn**” connection and hit “Test connection”.



From the IAM Role dropdown select IAM role created by CloudFormation template at the beginning.

IAM role name will be like - **RedshiftLab-GlueExecutionrole-<randomstring>** and hit Test connection.



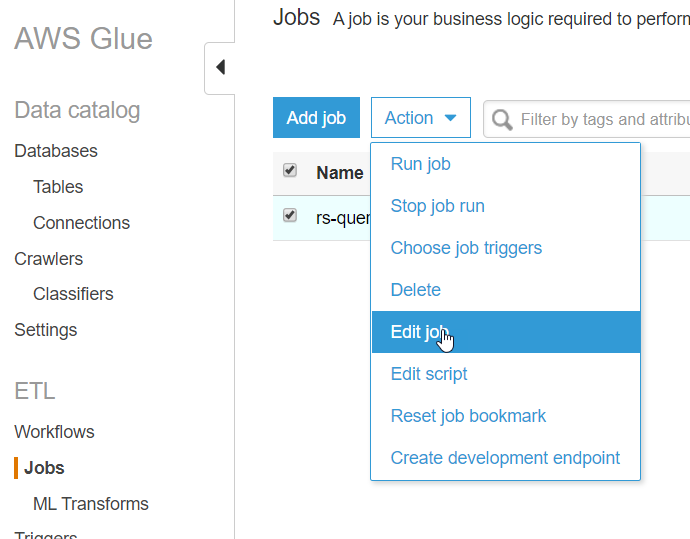
Note: It will take a couple of minutes to test the connection.

You will get a confirmation saying “**redshift-con** connected successfully to your instance”.

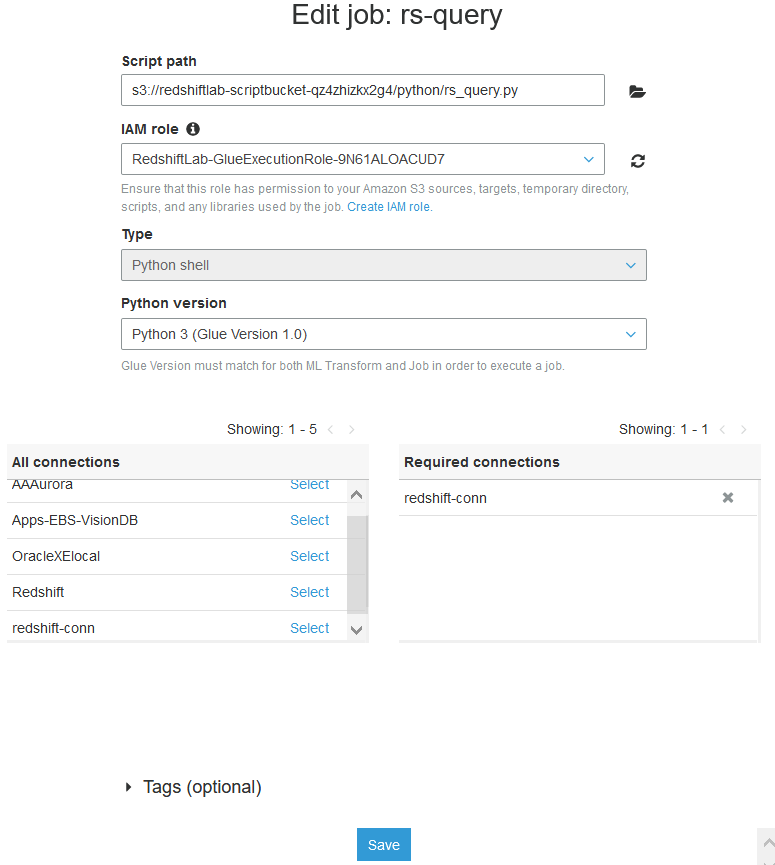
**Glue ETL Job**

Once connection is established you will update the Glue Job to include the connection information.

From the left pane **ETL > Jobs** Select the job“rs-query”and then **Action > Edit job.**



In this screen select the “**redshift-con**” connection from All connections so that it appears under Required connections. Hit save.



# **Redshift schema setup**

We will setup the schema and create the database tables from Redshift Editor.

Navigate **Services > Redshift**. From the left pane click **Editor**. If you have already connected to the editor in the section **Navigate to your Amazon Redshift cluster** then you will be automatically connected to the database. Otherwise follow the steps in the section **Navigate to your Amazon Redshift cluster**.

In the Query editor click on the “+” sign to open a new tab. Paste below scripts.

CREATE SCHEMA supplychainschema;

CREATE EXTERNAL SCHEMA spectrum\_schema FROM DATA CATALOG database 'default' region '{<enter-aws-region>}' iam\_role '{<enter ARN of IAM ROLE associated in Redshift>}';

CREATE TABLE supplychainschema.nyc\_greentaxi(

vendorid varchar(10),

lpep\_pickup\_datetime timestamp,

lpep\_dropoff\_datetime timestamp,

store\_and\_fwd\_flag char(1),

ratecodeid int,

pulocationid int,

dolocationid int,

passenger\_count int,

trip\_distance decimal(8,2),

fare\_amount decimal(8,2),

extra decimal(8,2),

mta\_tax decimal(8,2),

tip\_amount decimal(8,2),

tolls\_amount decimal(8,2),

ehail\_fee varchar(100),

improvement\_surcharge decimal(8,2),

total\_amount decimal(8,2),

payment\_type varchar(10),

trip\_type varchar(10),

congestion\_surcharge decimal(8,2))

;

You can find your IAM ARN by navigating to Redshift CLUSTERS > Properties tab

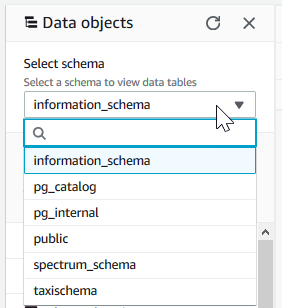


Update Screenshot



Execute each statement one after another by selecting the statement and hit Run.

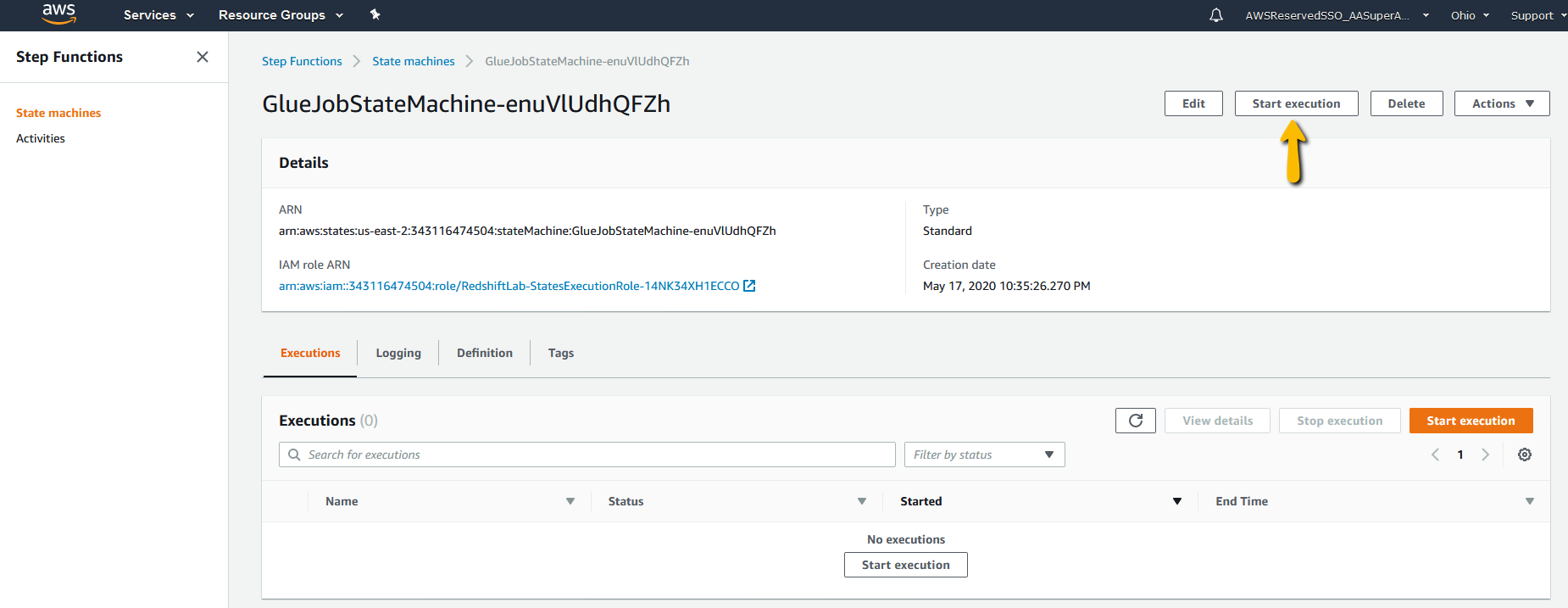
Once the statements are executed check the schema that got created in Redshift.



# ELT Execution

Now the State machine needs to be executed to orchestrate 1) ingestion into Redshift i.e **EL**(T) and 2) Unload the data into data lake i.e (EL)**T**

Navigate to **Step Functions** page and click on theState machineand hit Start execution. Leave the JSON input as is. Monitor the progress of the job under Visual workflow.



# Reporting

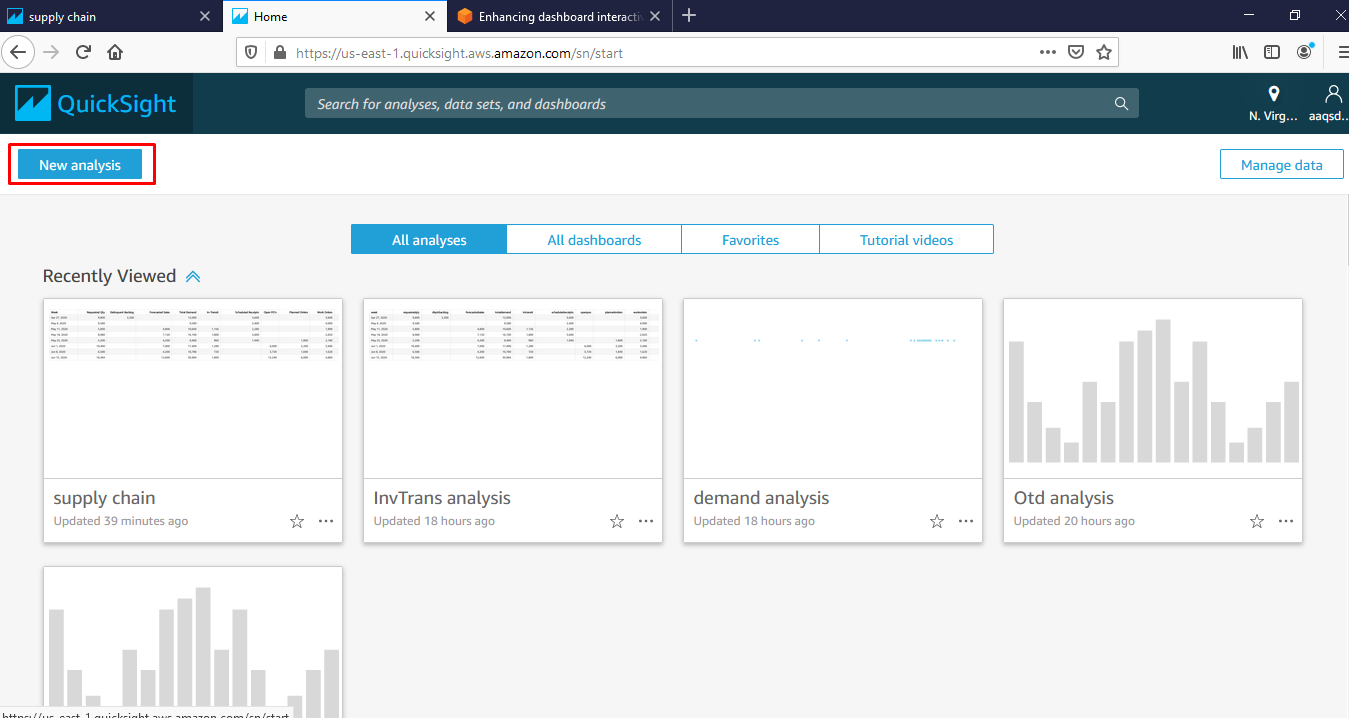
The above step brings us to for the first time to look at the data we just loaded.

Open **Quicksight** in AWS Console the Services options and then select the region as instructed by your instructor.

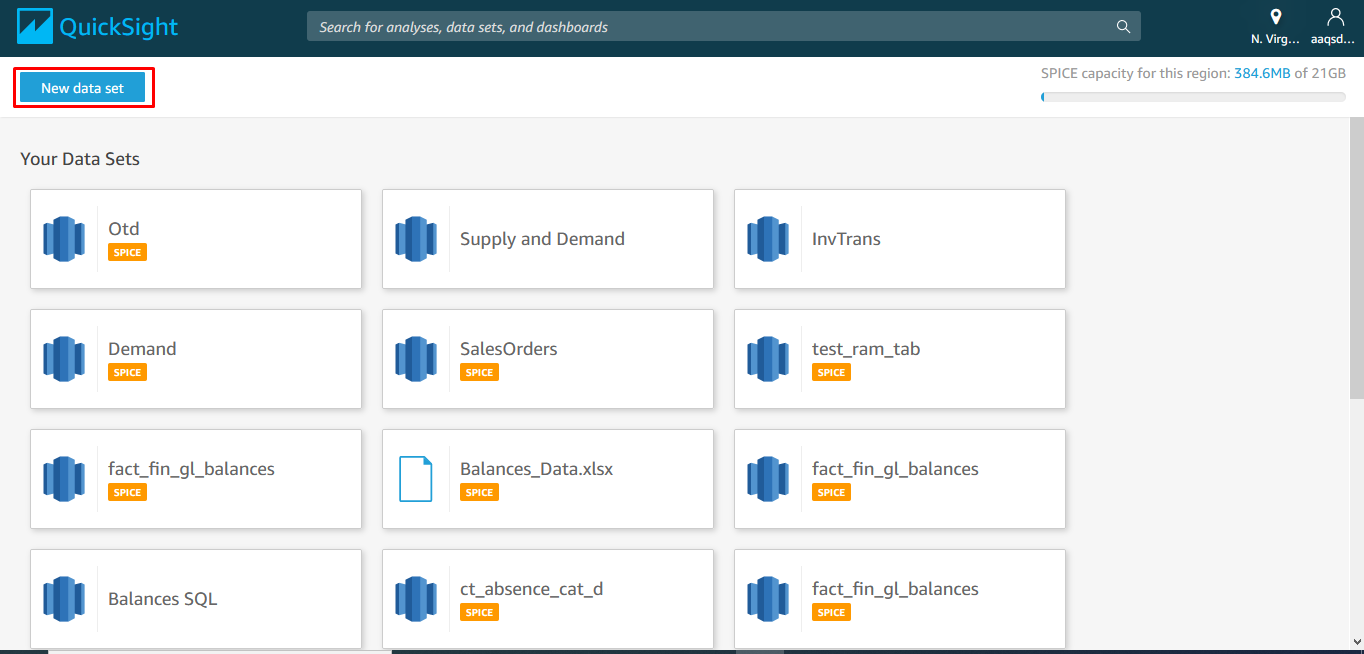
**Note**: AWS Credentials should have been shared earlier as part of the Immersion day. Please check if you haven’t received.

1.Connecting to Database to get Datasets

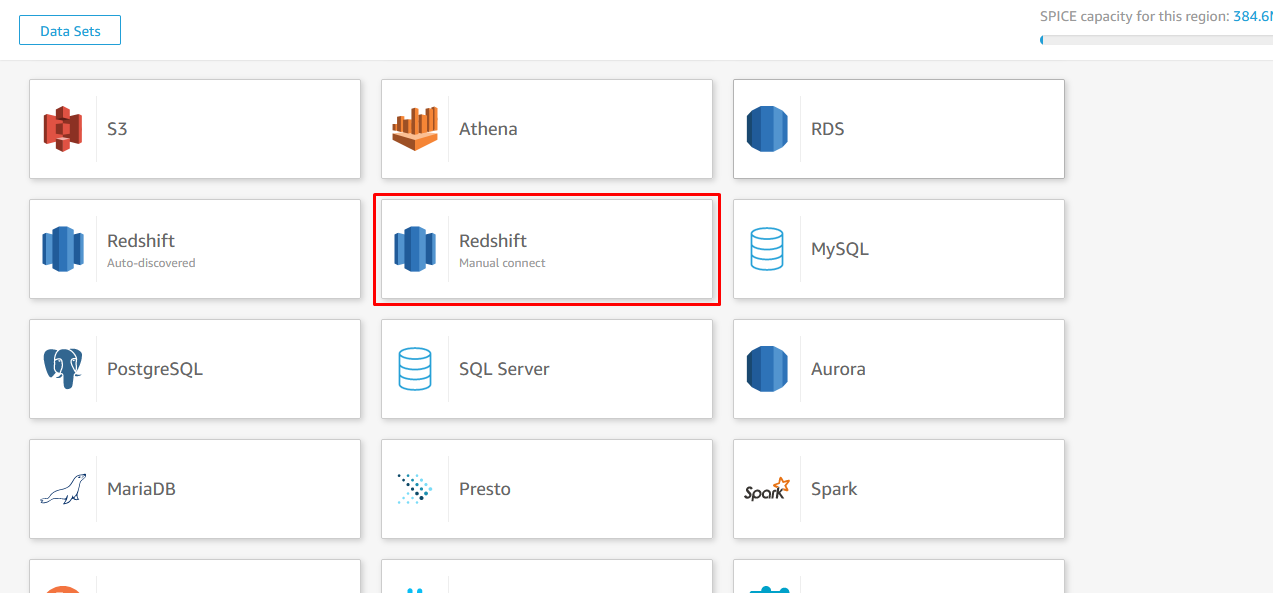
* Open Quick sight page. After logging in click on New Analysis



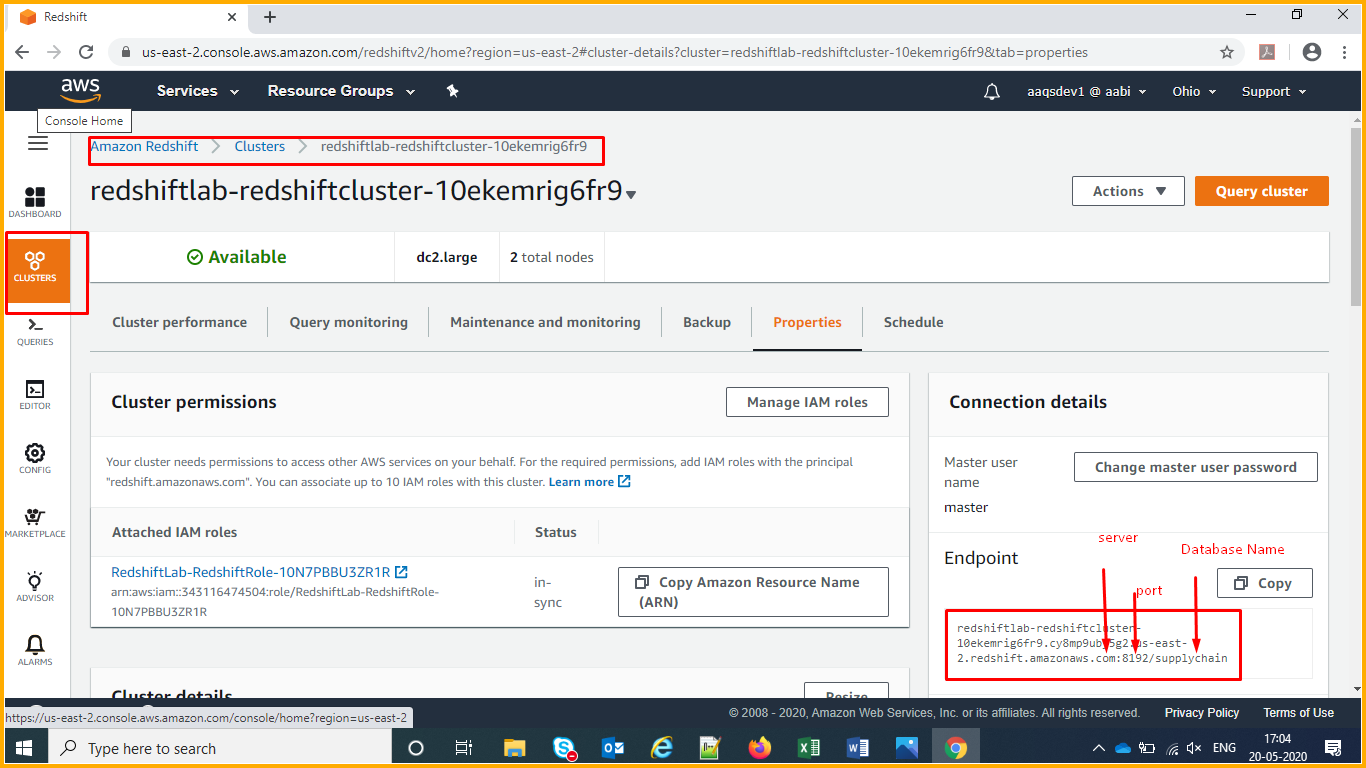
* Click on New Dataset



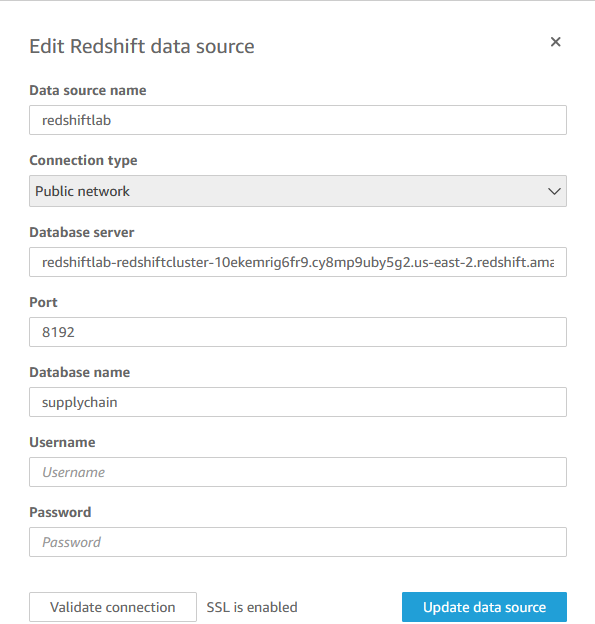
* Click on Redshift to get connected to Redshift DB



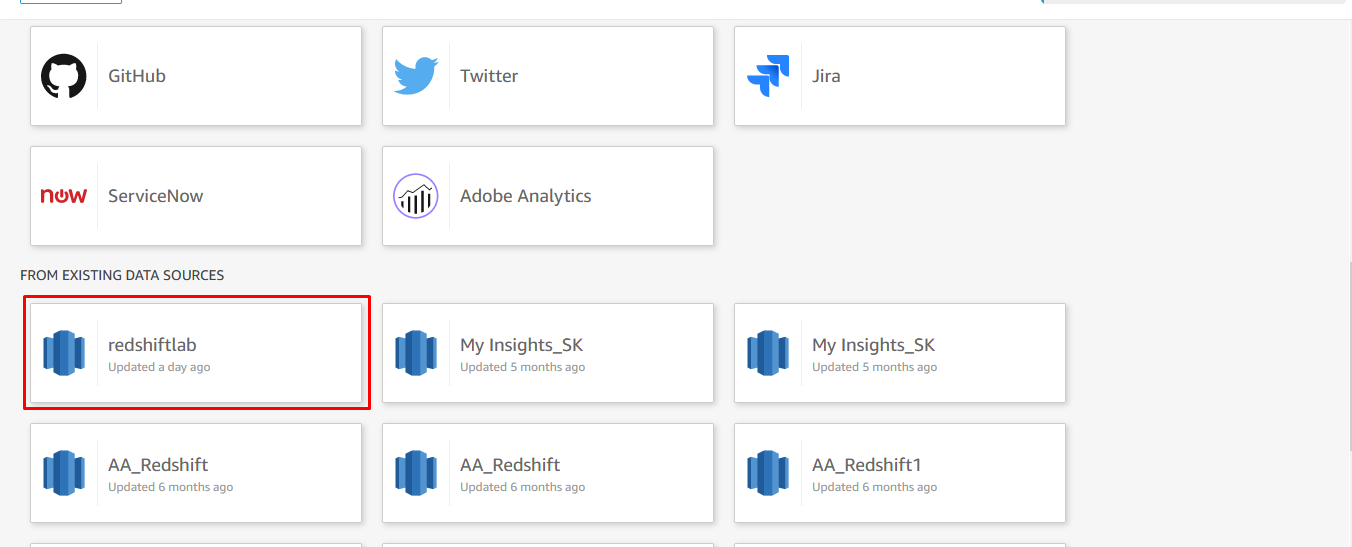
* Give Credentials required as shown. Database server and port can be found at



After finding the required details, use them as per the below image.



Once connected you can see the connection like this

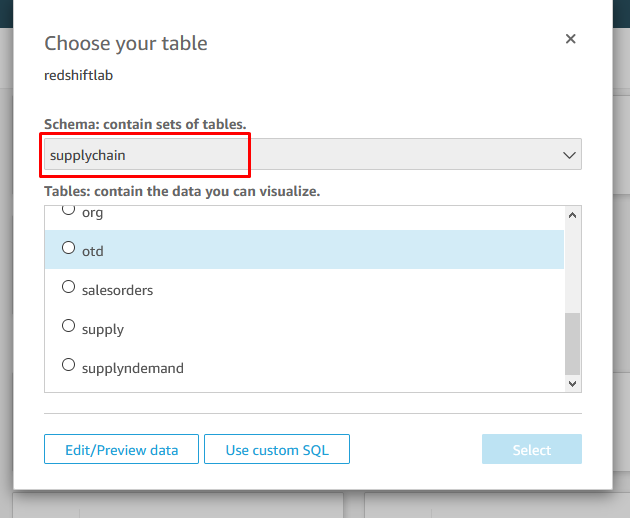


Once the datasource connection is done.

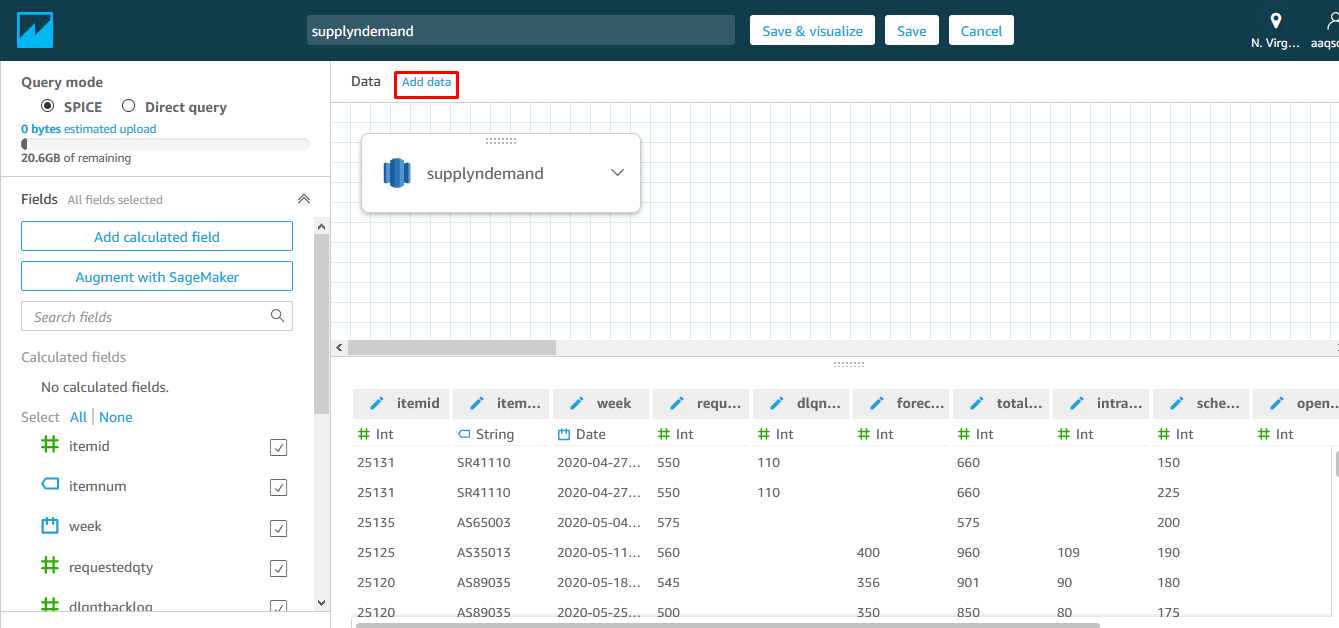
Click on the Data Source highlighted above.

we can create Dataset as shown below

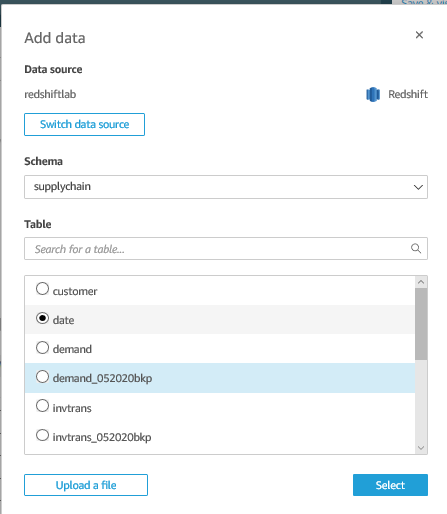




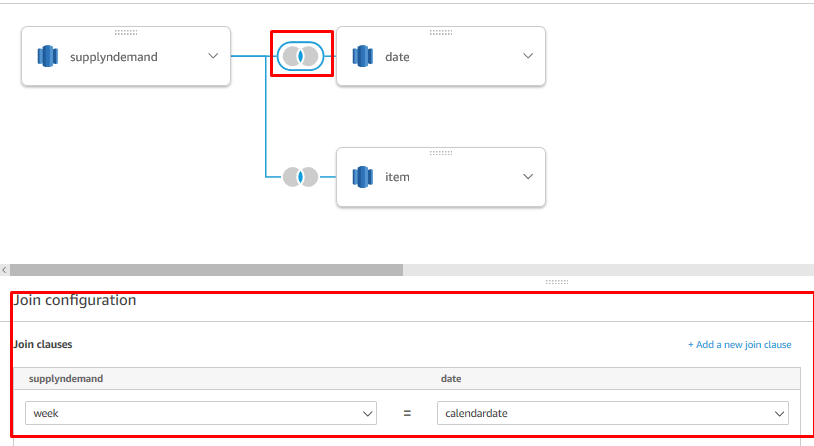
After selecting supplyndemand Table click on Edit/Preview data to add more tables for creating the required dataset. Click on Add Data to add tables as shown below.



Now select date table and click on Select



Select the join option and give the join as shown.



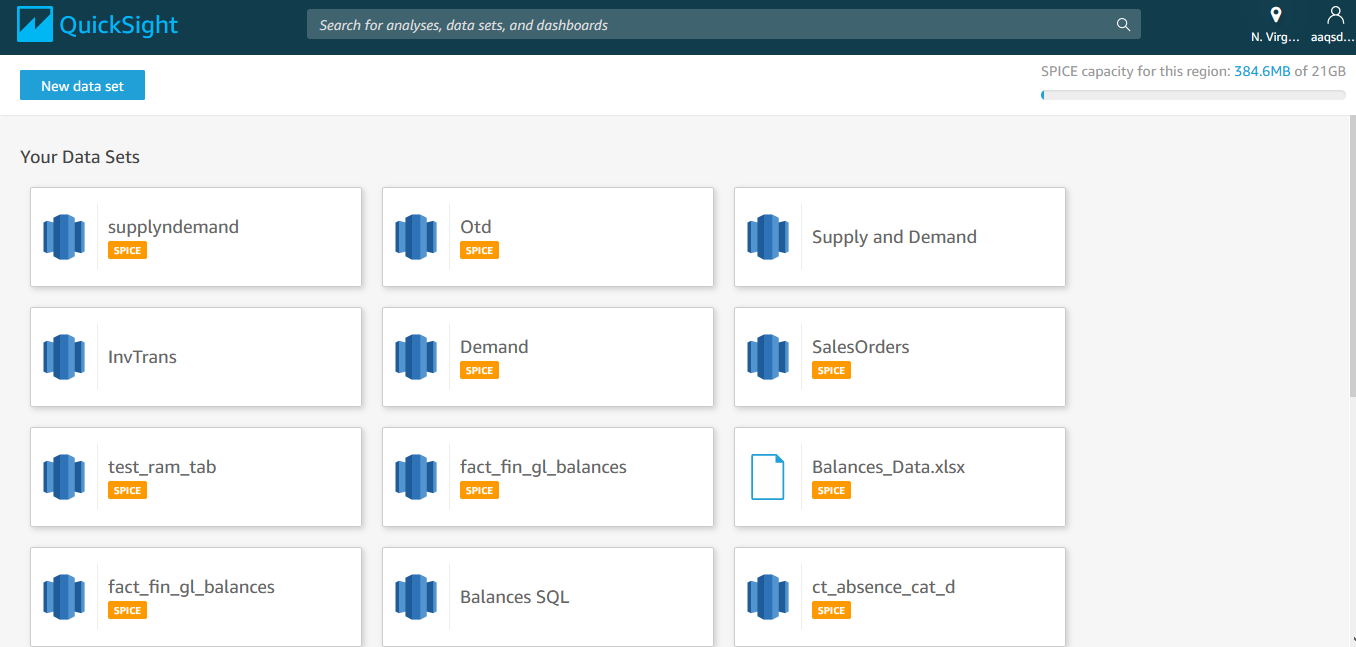
Similarly drag all the tables and give the joins as shown below to create the dataset for Supplyndemand

a)Supplyndemand

Tables and joins

Supplyndemand(ItemId) – Item(ItemId)

Supplyndemand(week) – date(calendardate)



By following the above steps we can create the required datasets. Once we create all the required datasets, datasets can be seen as per the screen above.

The following datasets are required for creating the dashboard.

a)Supplyndemand

Table(joincolumn)

Supplyndemand(ItemId) – Item(ItemId)

Supplyndemand(week) – date(calendardate)

b)OTD

Table(joincolumn)

Otd(item id) - Item(item id)

Otd(week) - date(calendar date)

Otd(customernum) - Customer(customernum)

c)Salesorders

Table(joincolumn)

SalesOrders(item id) – item(item id)

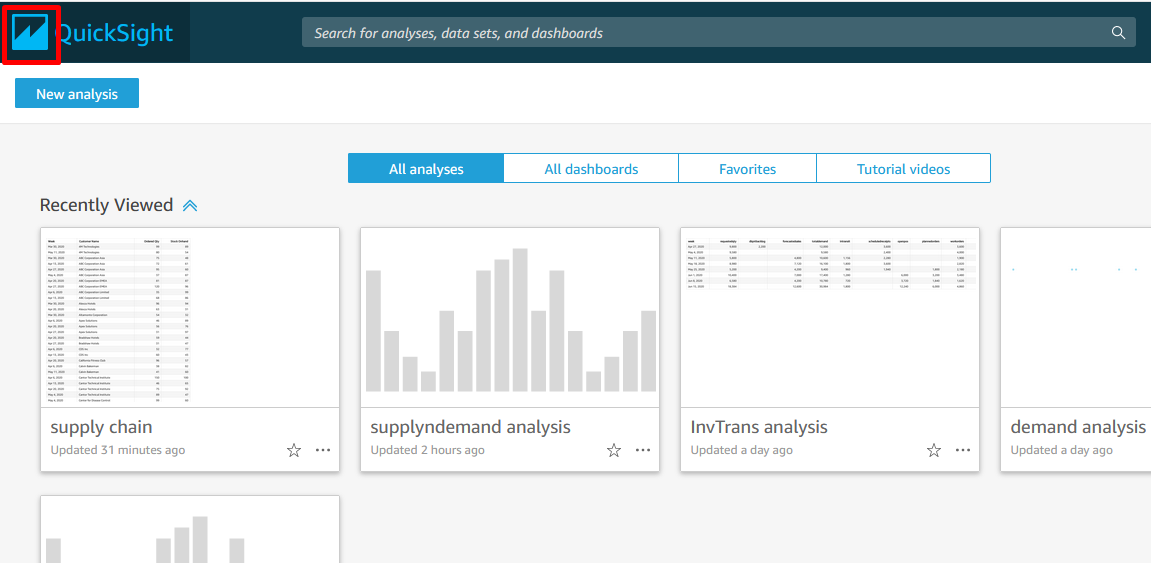
SalesOrders(week) – date(calendar date)

SalesOrders(customernum) - Customer(customernum)

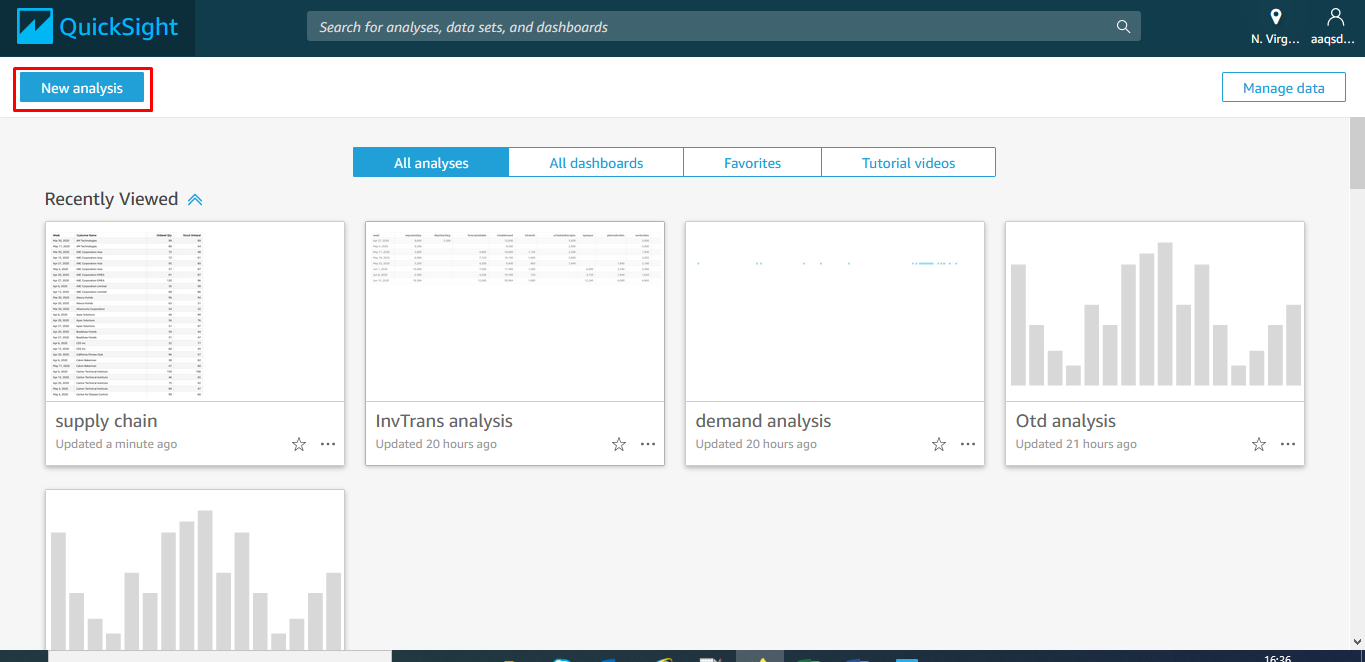
d)InvTrans(item id)—item(item id)

2.Creating Reports:

Click on the below highlighted part to navigate to home

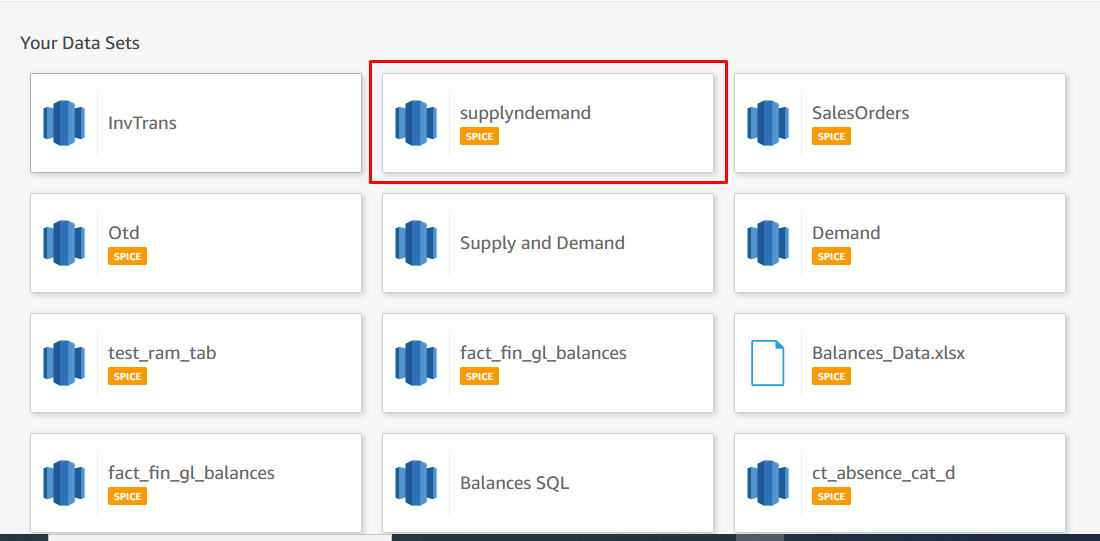


Click on New Analysis.

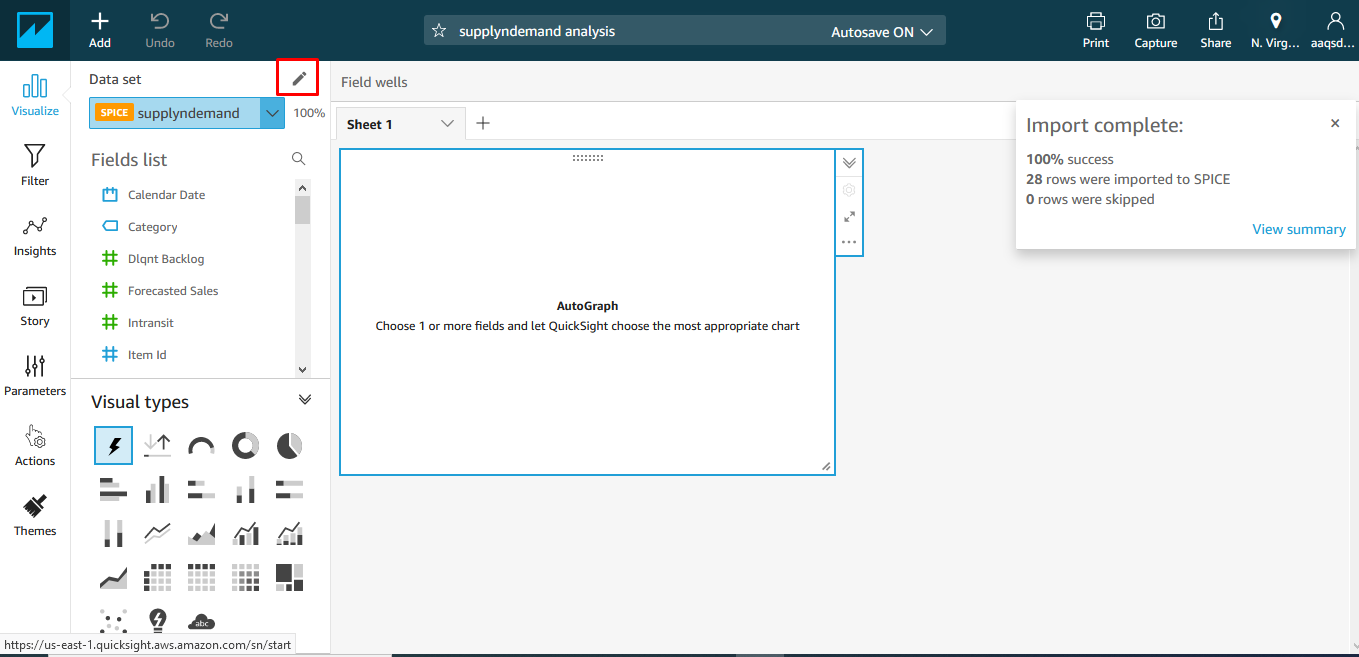


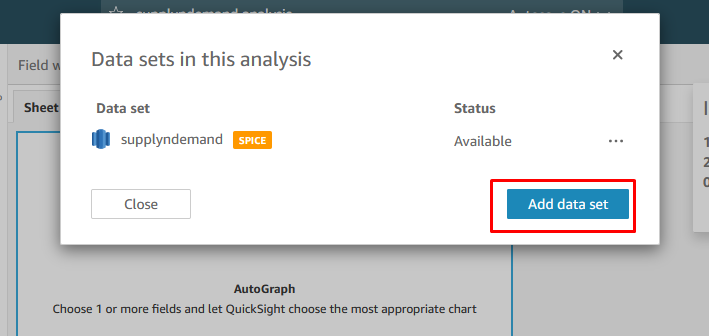
**Below points show a summary of creating the datasets and blank visual**

Click on the Dataset as shown below

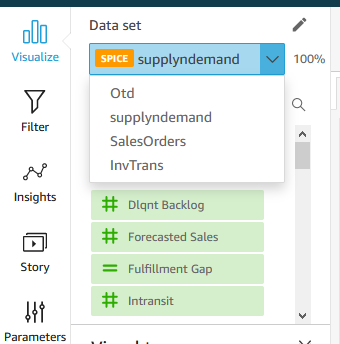


If we are using multiple datasets in a single dashboard, we need to bring them all. By Following the below step . Click on the edit symbol shown below.



Click on Add Dataset

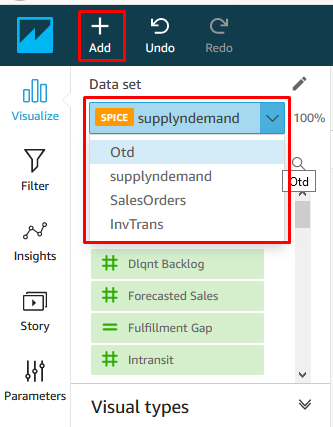
Add all the required datasets one after the other and finally after adding all the required datasets, we can see the one shown below



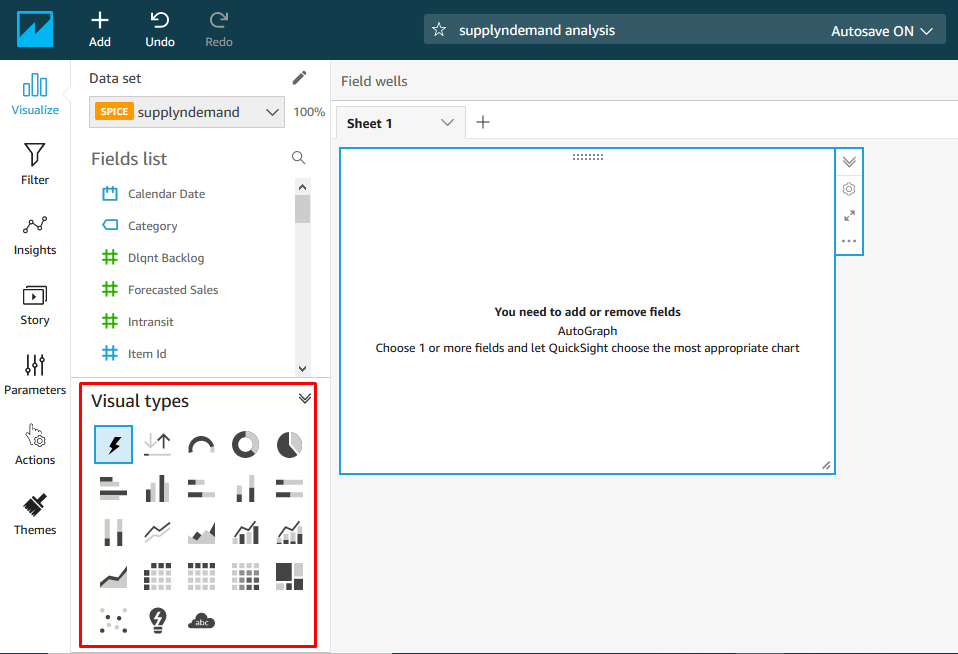
Once all the Datasets are added, we can start creating visual.

Note: We can use only one Dataset for creating a Visual.

Select the Dataset, you want to use for creating visual and click on add to create visual on the dataset.



After creating a new visual, by default it will be as shown below. Change the visual type as required



We now know how to import all the datasets and a blank visual. Let’s create all reports one by one

Product Categories Report:

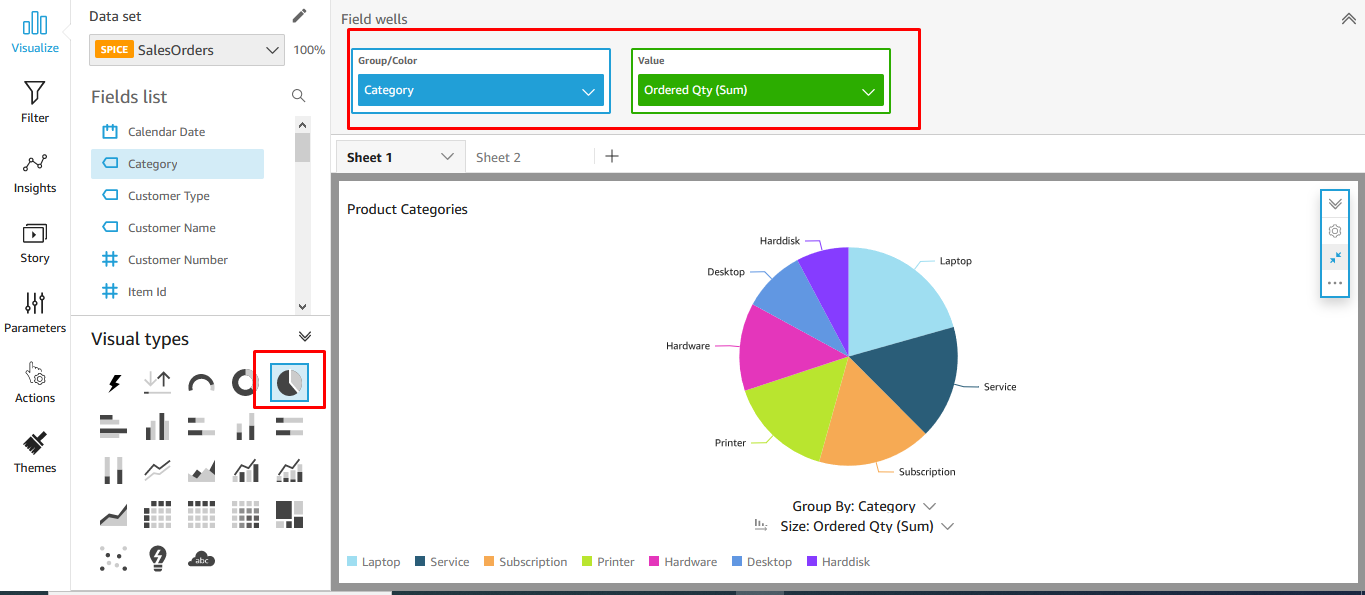
1.select sales order dataset from the drop down under datasets

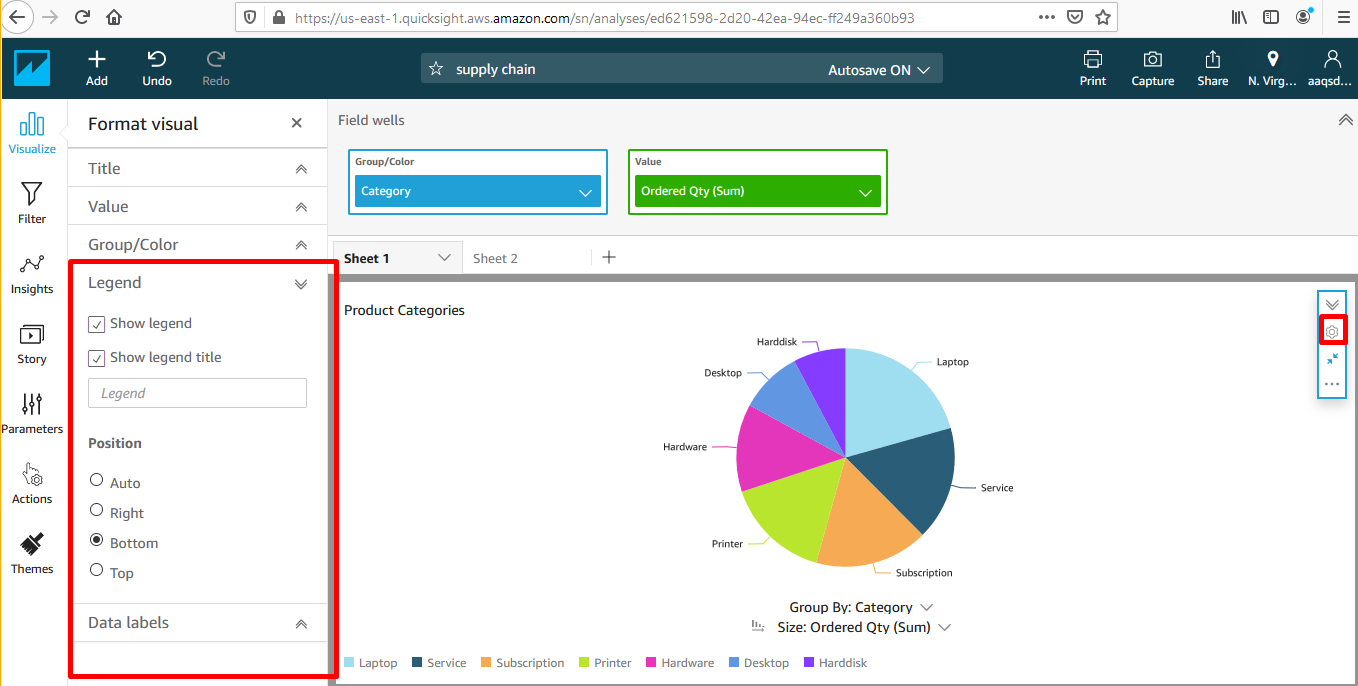
2.Add visual as shown above

3.Select the visual as pie from the visual types section as shown below

4.Drag the category into group/colour and Order Qty into value in the field wells section as highlighted

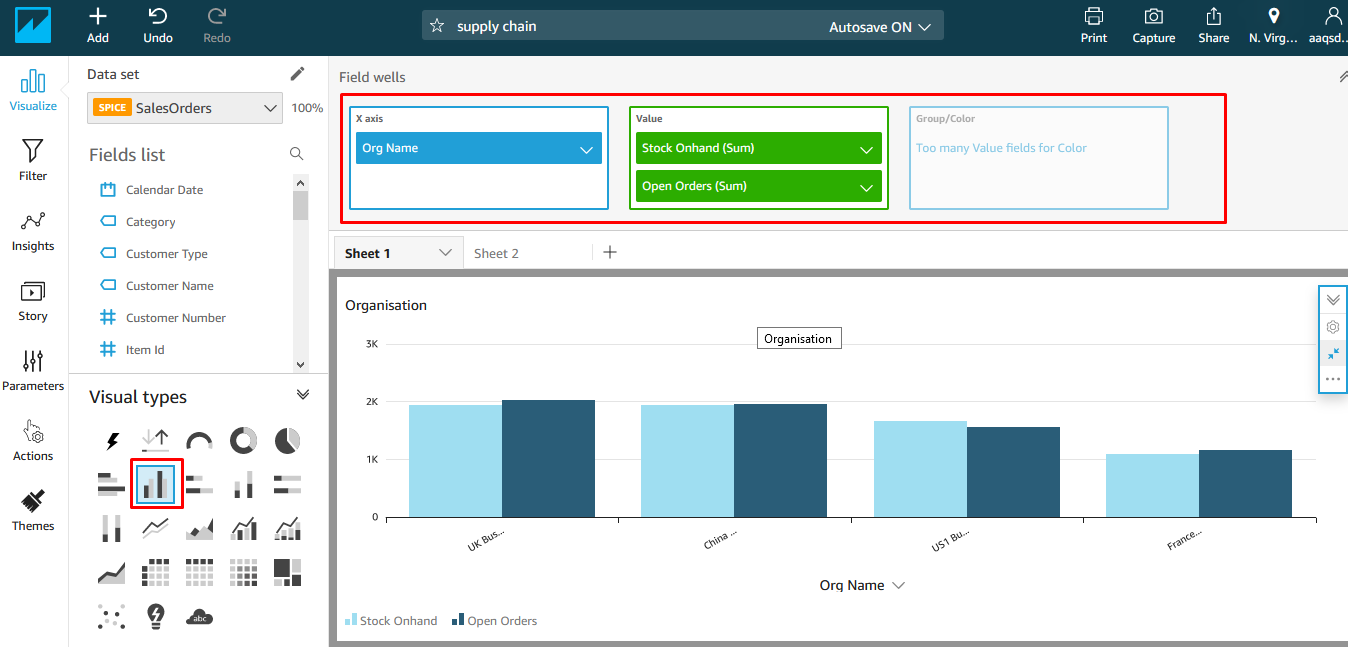
5.Click on the Title shown above the diagram to edit the title



6.Click on the Setting button highlighted below and Expand legend and select bottom as shown to place the legend at the bottom 

Organisation Report :

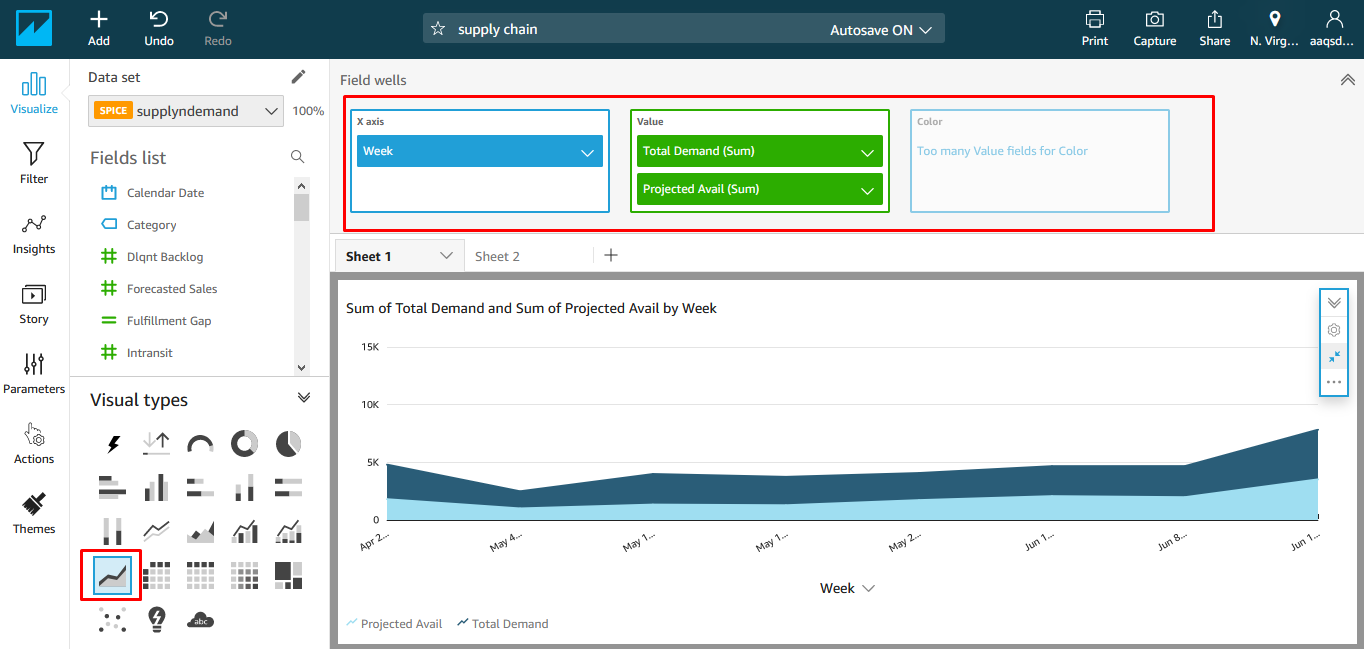
1.Select the dataset salesorders

2.Click on add new visual and create the report as shown below

Demand and supply by week

1.Select the dataset supplyndemand

2.Click on add new visual and create the report as shown below



Delays in fulfilling orders report

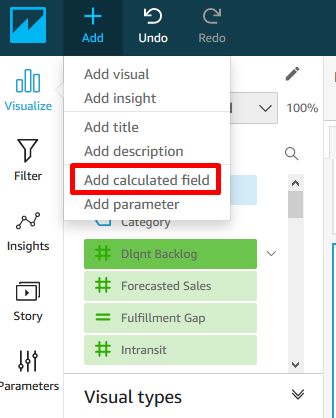
1.Select the dataset supplyndemand

2.Click on add new visual and create the report as shown below .

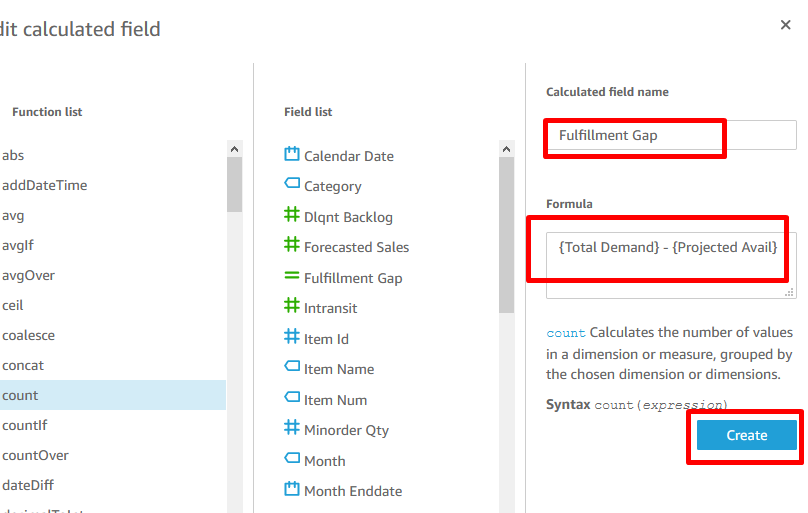
The fields to be added are Reqiested qty,delinquent banking,forecasted sales,total demand,intransit,scheduled receipts,open po’s,planned orders,work orders,project availability,fulfilment gap.

Fulfillment Gap is a calculated field created as shown below

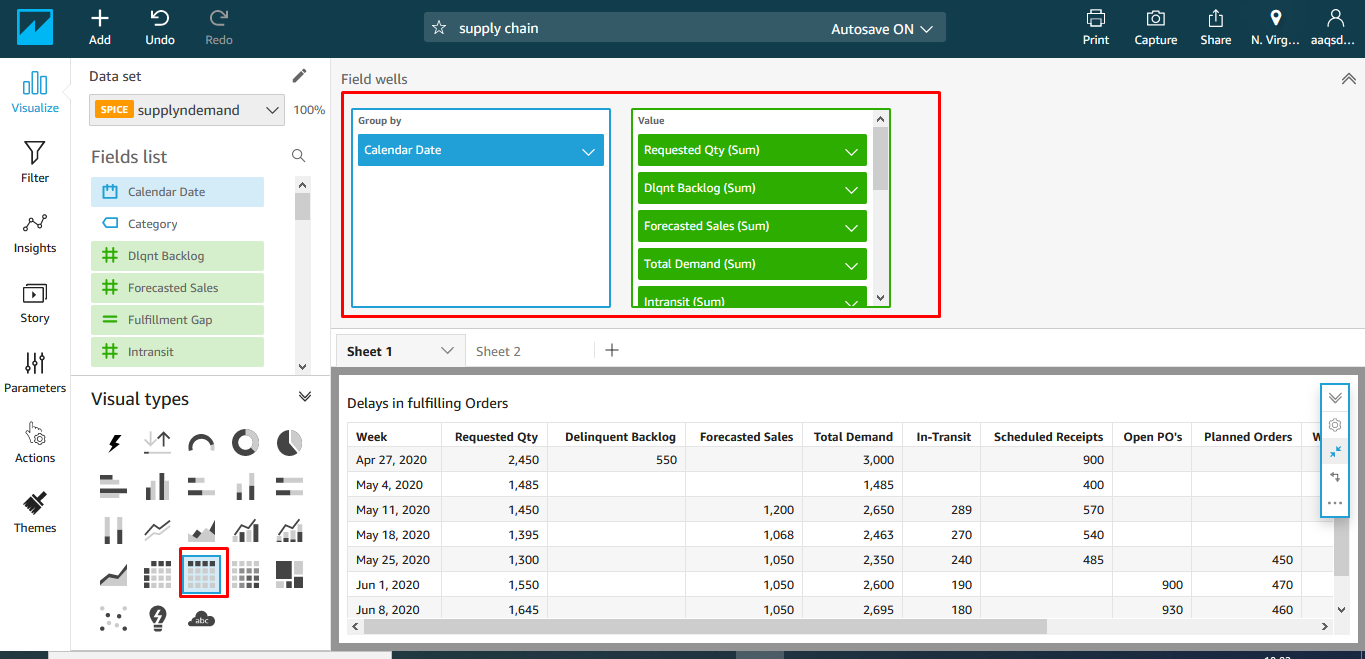
Select Dataset as supplyndemand ,click on add and select calculated field as shown below



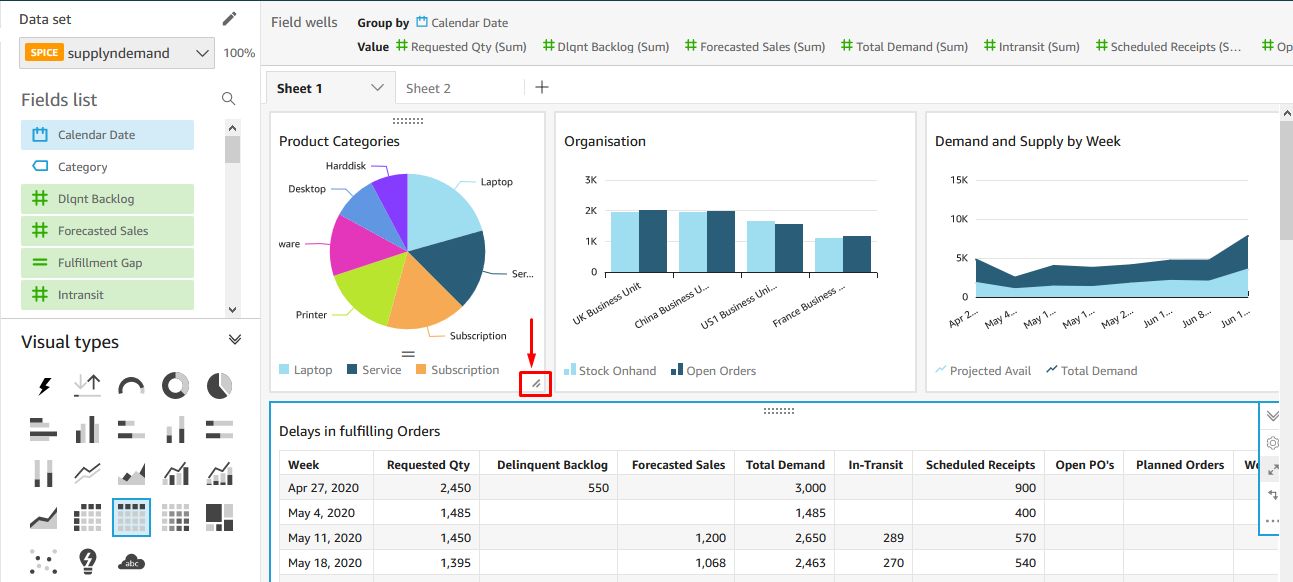
And then



Delay in fulfilling orders report is shown below

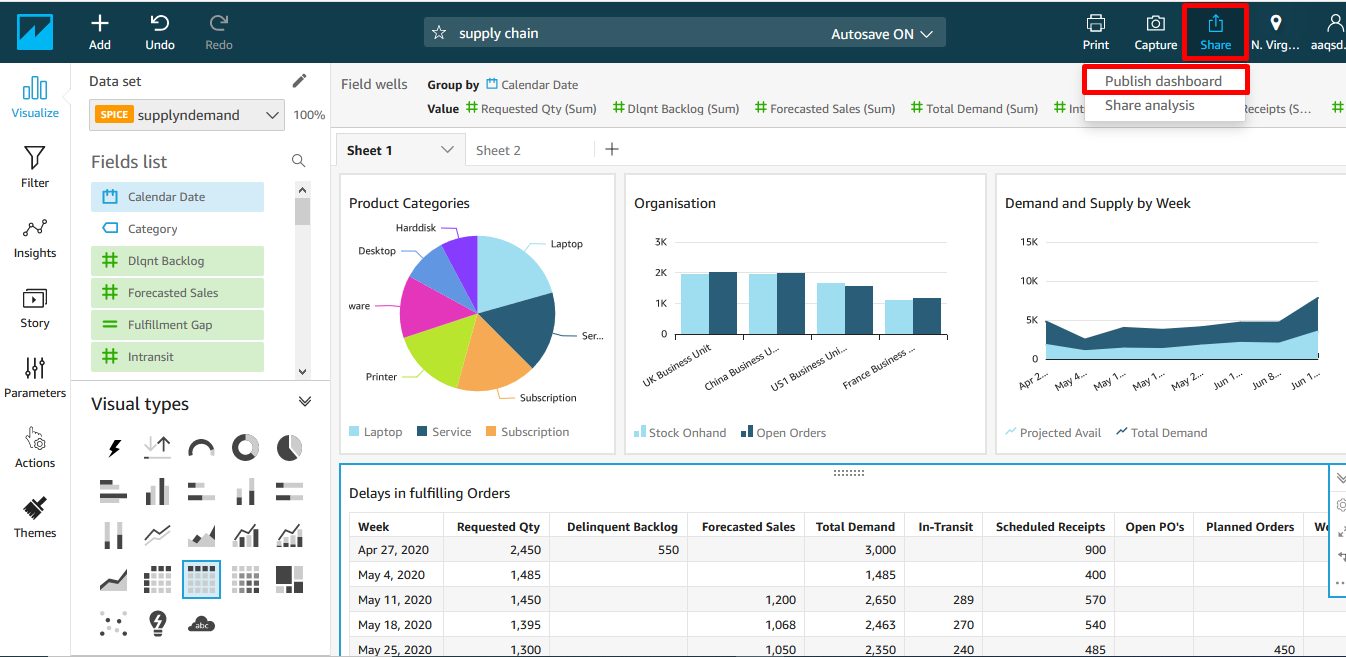


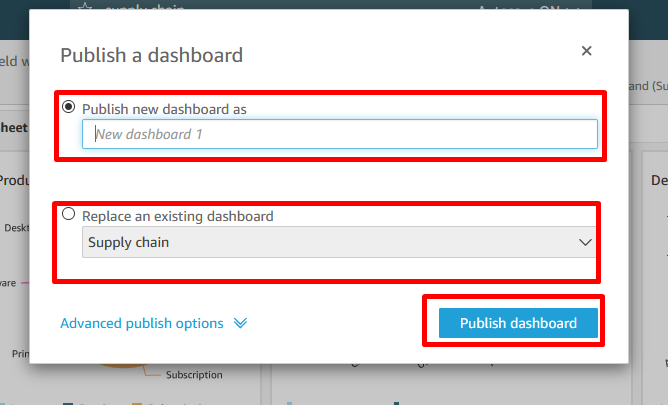
We can adjust the size of the report by dragging the button highlighted and arrage the visuals as shown below



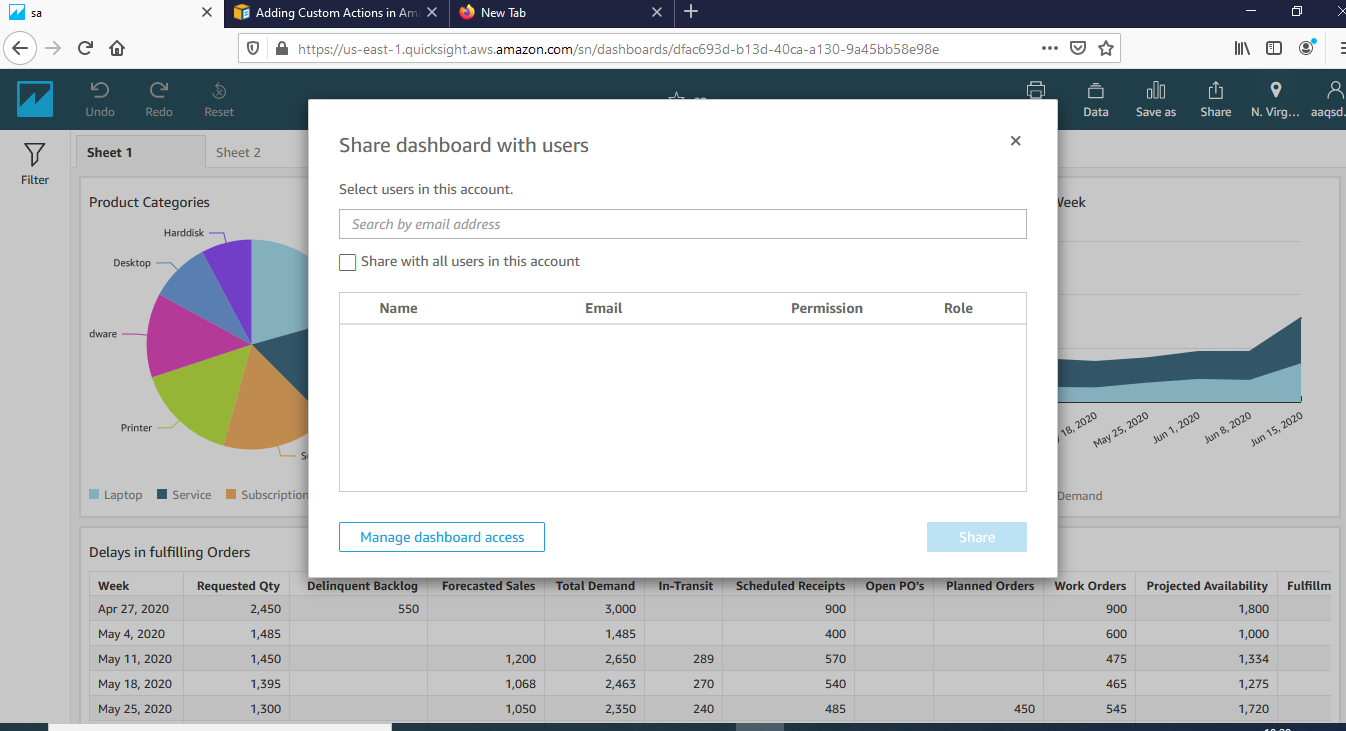
Once the reports are adjusted ,we can make this analysis into dashboard by publishing it.

Publishing Dashboard

Click on share,then click publish dashboard

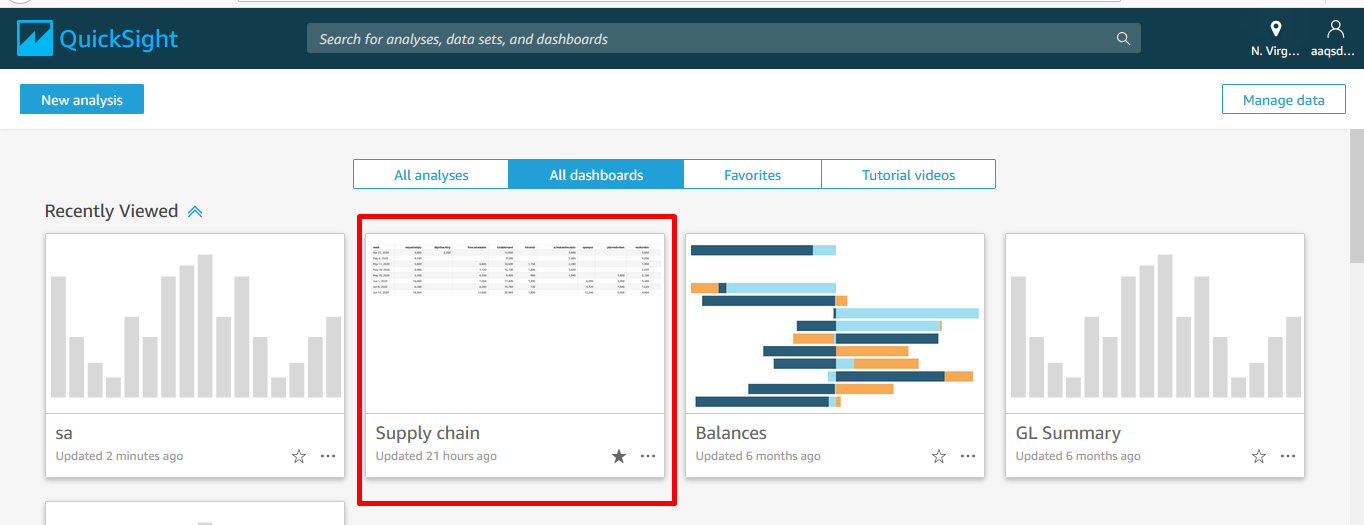
If we are publishing a new dashboard give a name and publish or if we are replacing the existing dashboard then select the existing dashboard and publish

Once you click on publish dashboard, the dashboard will be published and it will give a prompt as shown below.

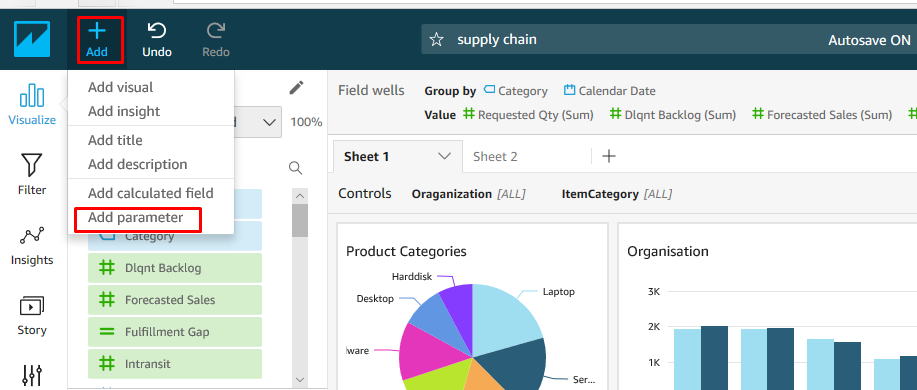


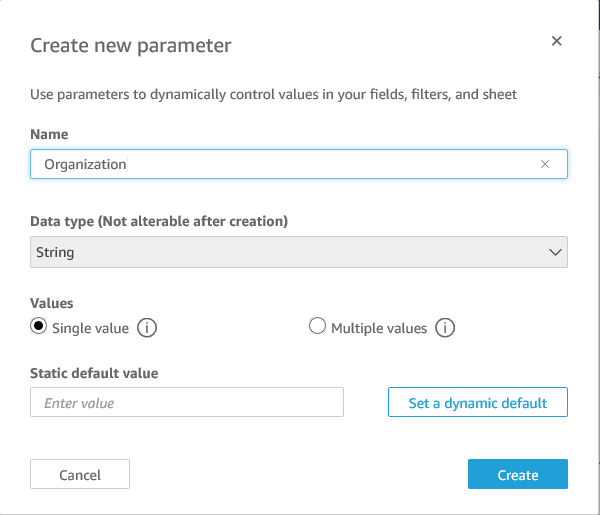
If you want to share it to a particular mail you can or you can just ignore it by closing it.

In the home page in all dashboards you can find the dashboard created.

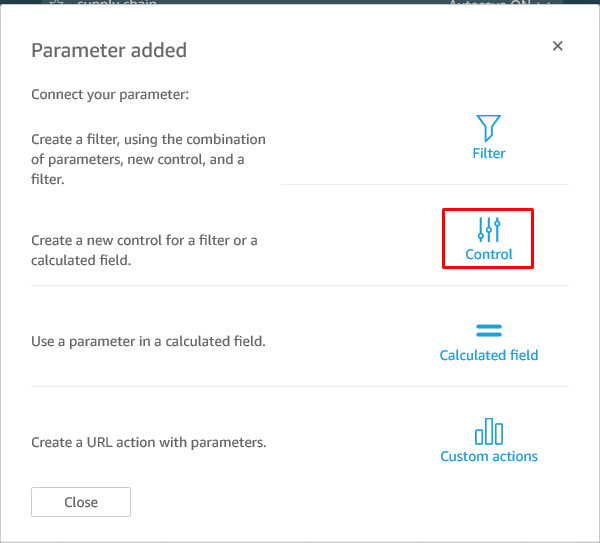


**Creation of Prompts:**

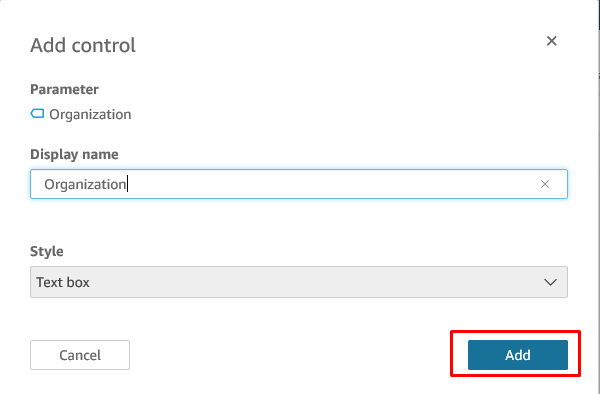
1. Prompts that work on single dataset
2. Prompts that work on all datasets
3. Prompts that work in Single dataset.
   1. Create this prompt if that column is not present in other datasets eg: Organization. Organization is present in Sales order dataset but not in others.
   2. Click on Add and Add parameter as shown.
   3. Then Click on Create after giving the details as shown below.



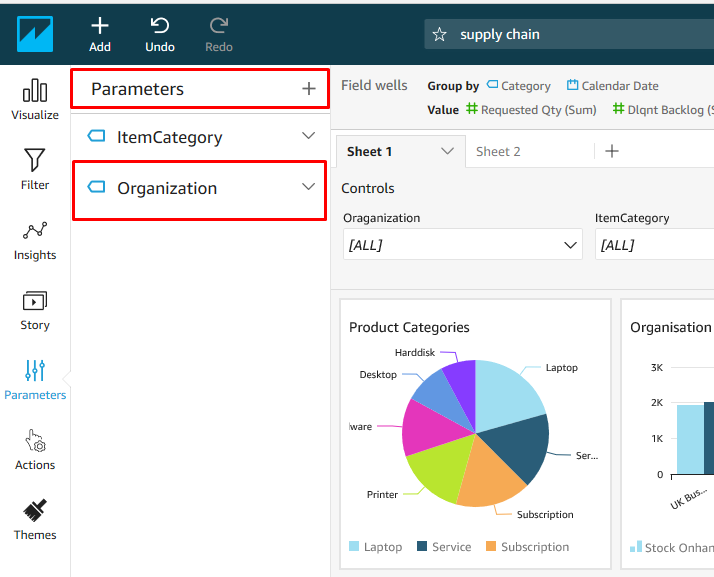
* 1. Click on Control in the next screen as you see below.



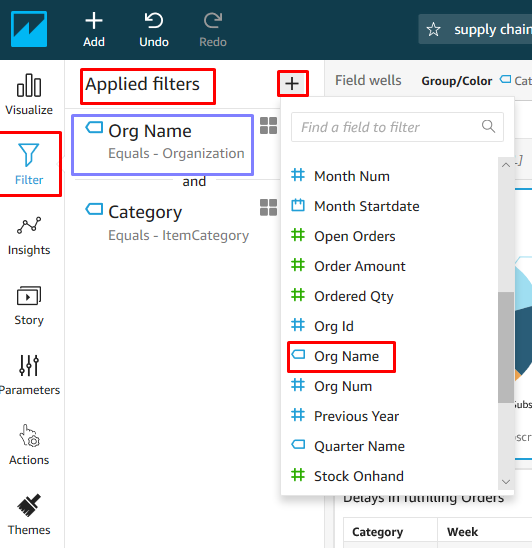
* 1. Click on Add after giving the display name as below.



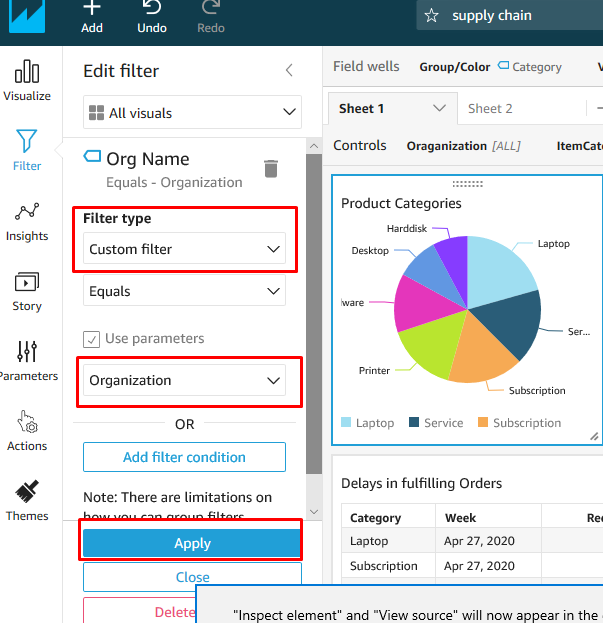
* 1. Then the added parameter is shown as below under parameters section.



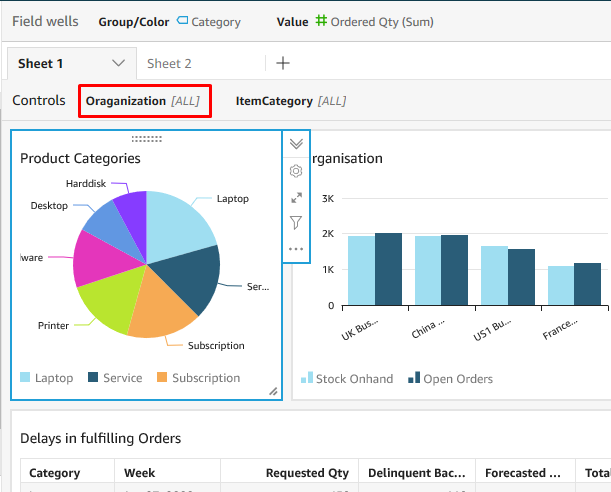
* 1. Click on the Product Categories Visual and select Filter section as highlighted and click on + icon beside Applied filters and select the Org Name. After selecting it Org Name is displayed as highlighted in blue.



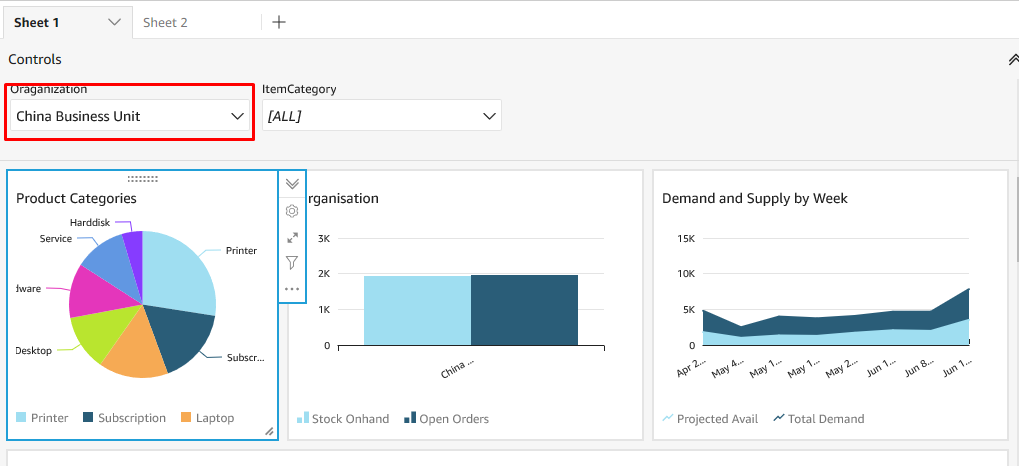
* 1. Click on the Org Name which is highlighted in blue before and it will be expanded. Choose the ones highlighted and click on Apply.



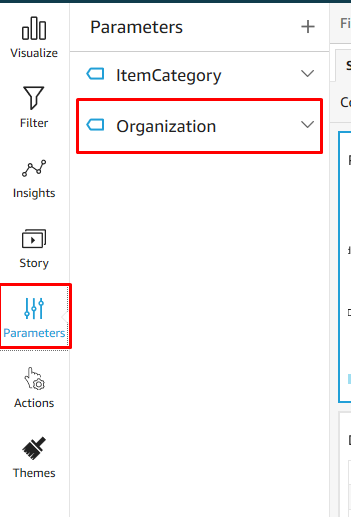
* 1. Then Organization Control will be displayed above all reports as shown. Click on prompt to expand.



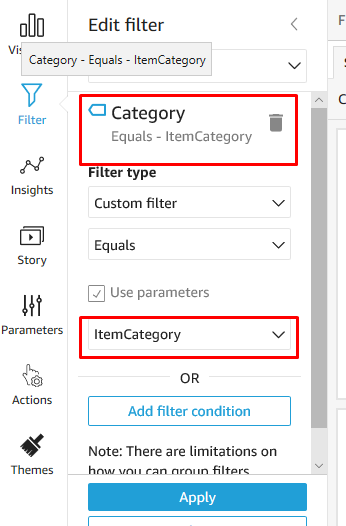
* 1. Select a value in organization and only the reports created on SalesOrders dataset is effected and Demand and supply by week report is not effected.



1. Prompts that work on all datasets
   1. For Eg., ItemCategory is present in all datasets, which can affect all the reports if we can create a prompt on it. Follow the same steps to create a Parameter called Category, that should display as shown below. Make sure that you select Parameters as highlighted below.



* 1. Select the Demand and Supply by week visual and create a filter on Category as shown below. We need to select the parameter we have created before.

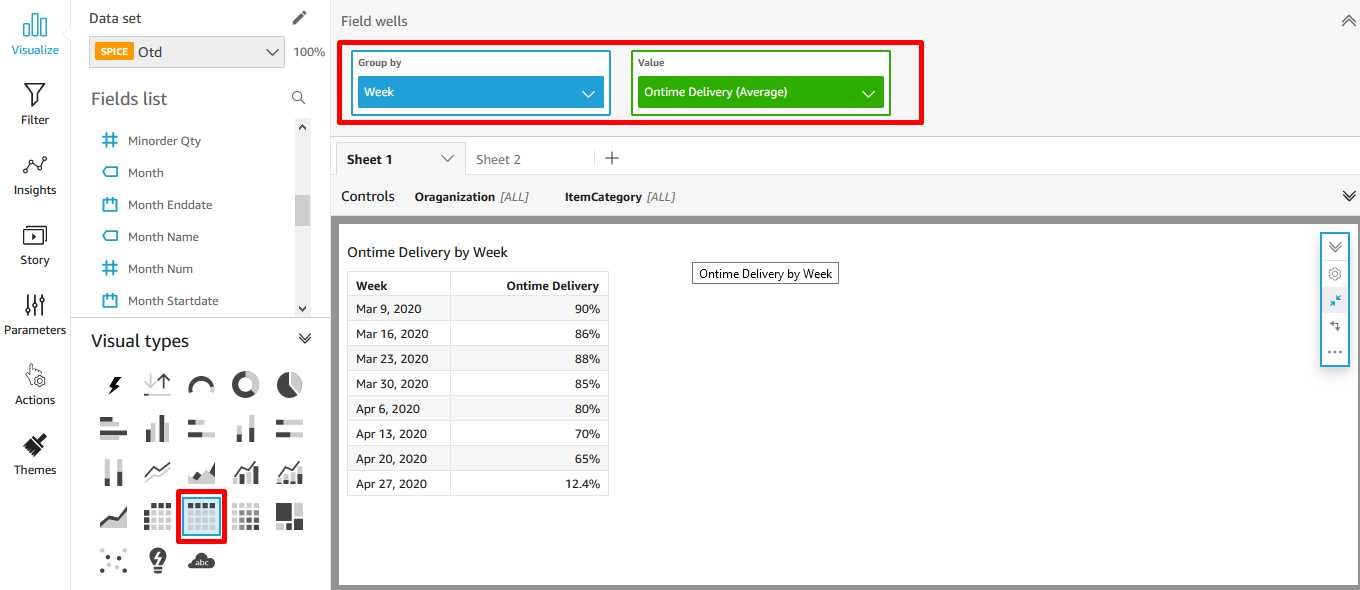


* 1. Also add a filter by selecting the visual product Categories and create the same filter as above. Then two datasets will be affected. Similar we can create for all the different datasets that we have.

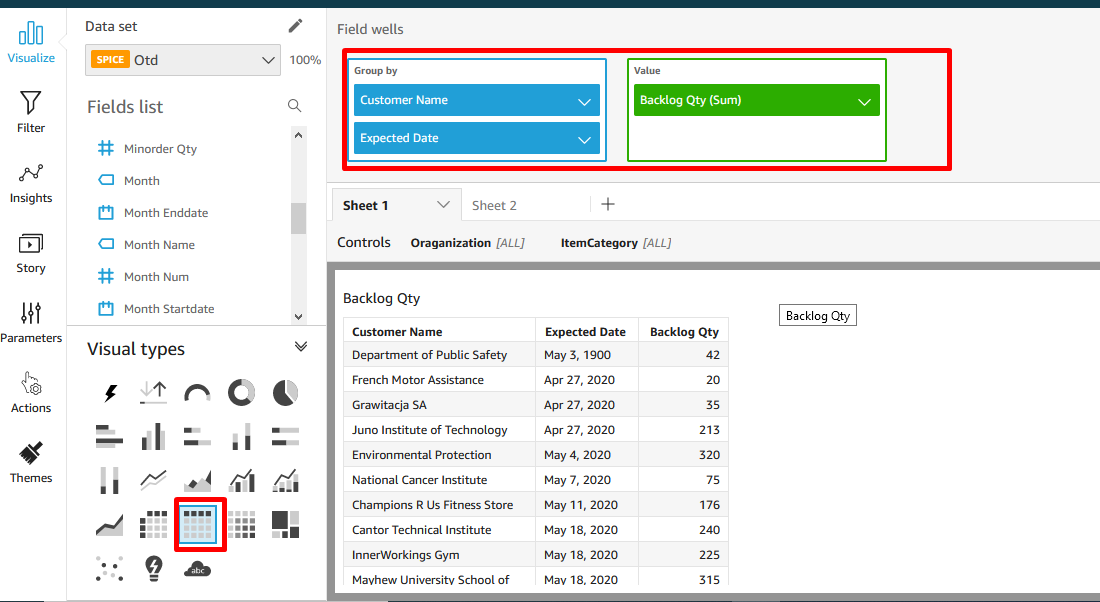
**Once all the reports are created publish the dashboard again.**

**Remaining Reports:**

Ontime Delivery Week

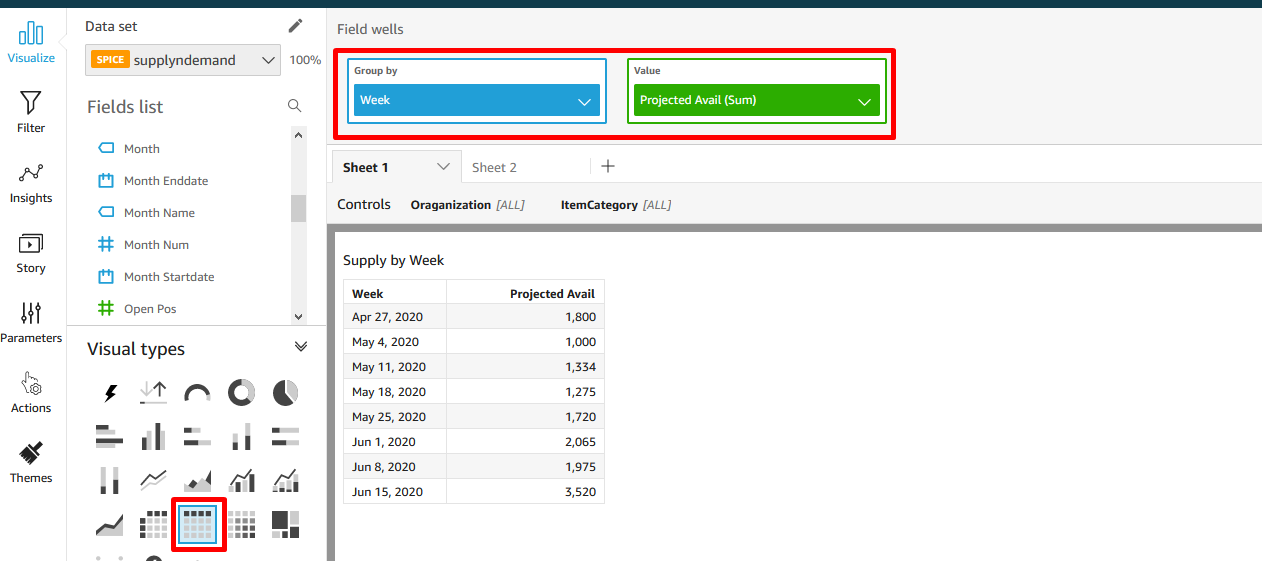


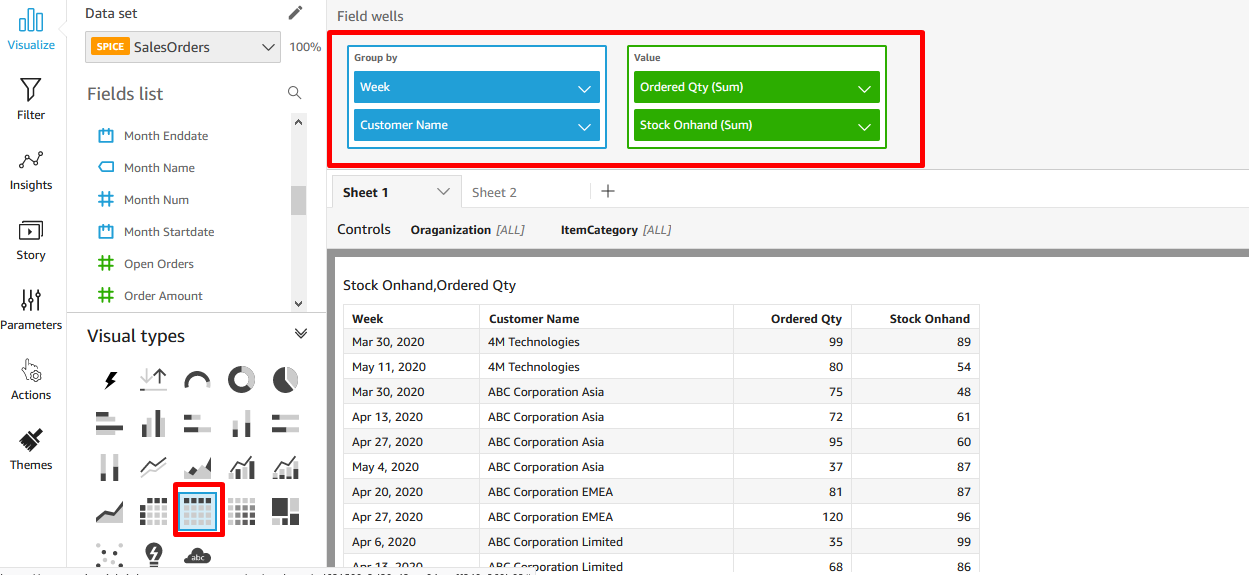
Backlog Qty



OTD % Report

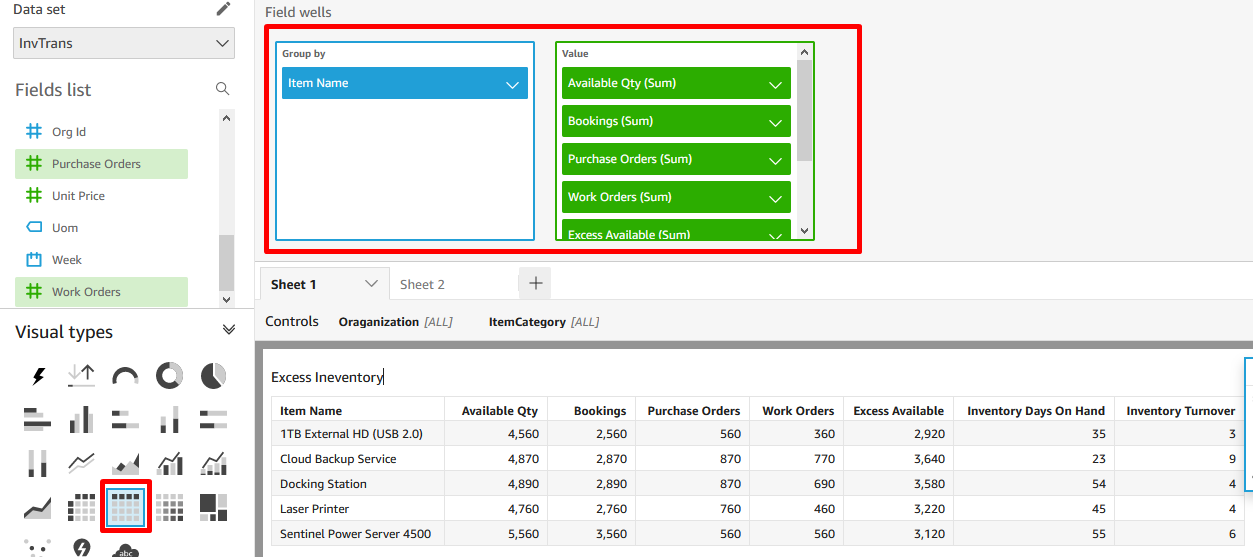
Supply by week report



Stock Onhand,Ordered Qty

Excess Inventroy Report

The Columns used in the report are Item name,AvailableQty,Bookings,Purchase orders,Work Orders, Excess Available.Inventory Days On Hand,Inventory Turnover



# Delete AWS resources from Cloud formation console.

Select the stack you created at the begin and click DELETE Stack.

Please fill in the survey at the end of this lab.