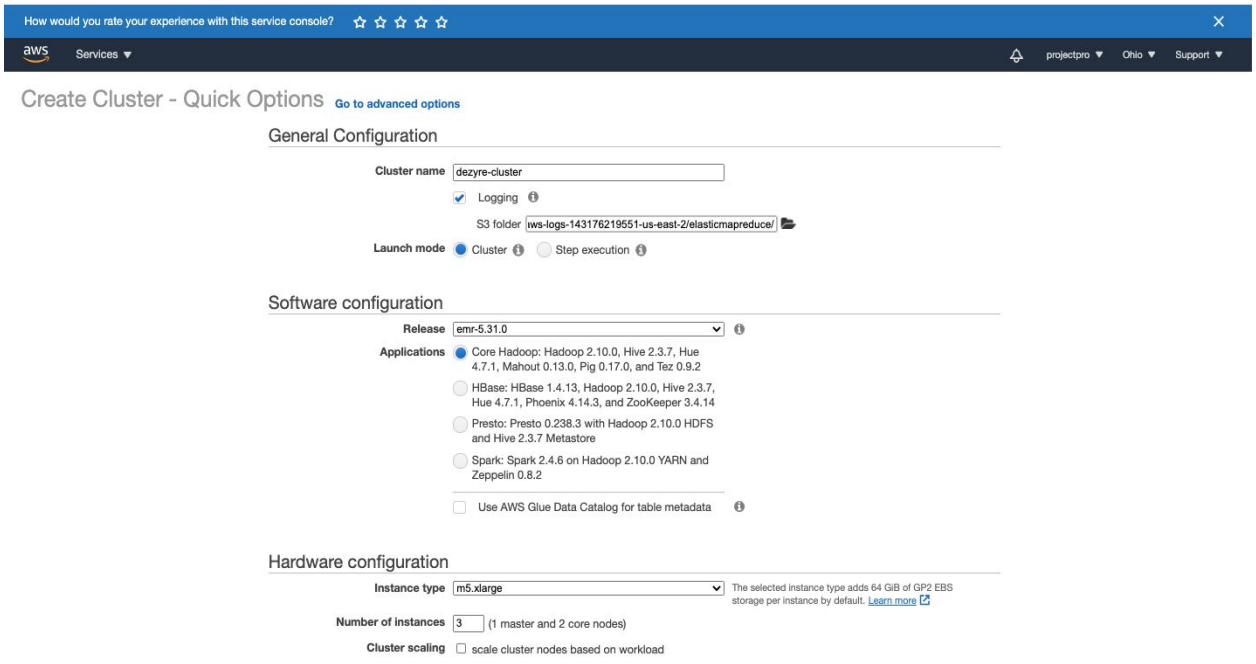
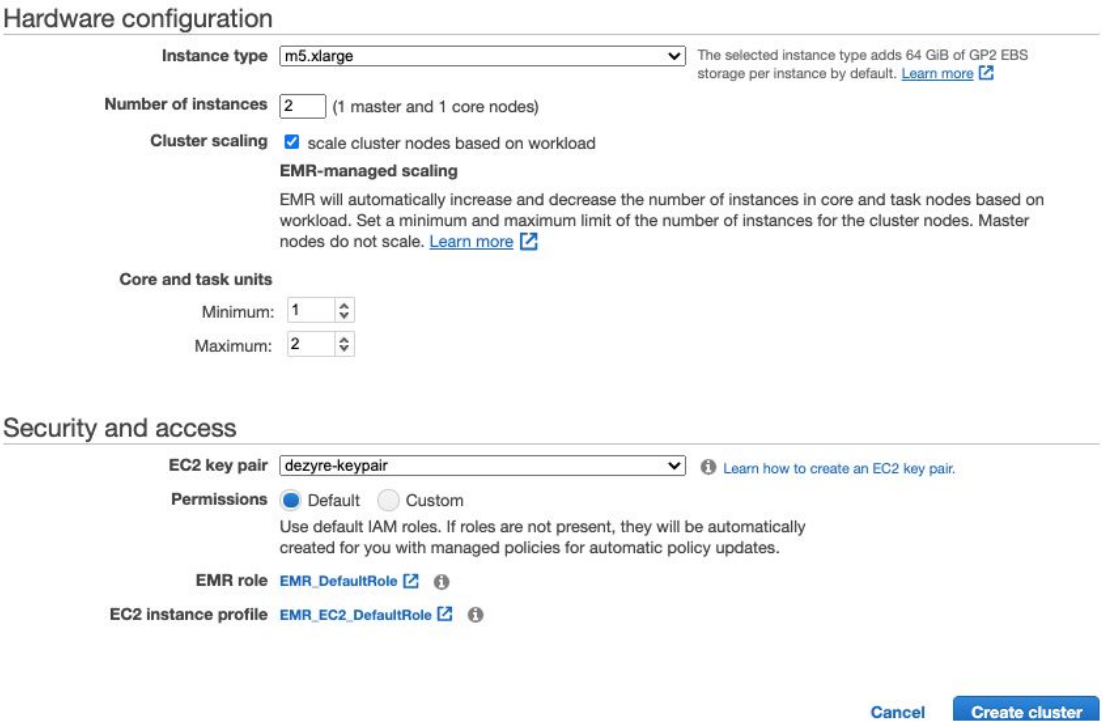
Project execution documentation

Step1 : Create EMR cluster in aws

1. Sign in to the AWS Management Console and open the Amazon EMR console
2. Choose Create cluster.
3. On the Create Cluster - Quick Options page, accept the default values except for the following fields:
    a. Enter a Cluster name that helps you identify the cluster, for example, dezyre-cluster
    b. Under Security and access, choose the EC2 key pair that you created in Create an Amazon EC2 Key Pair.
    c. In the applications, choose your required services , for our project we choose the first option
    d. In hardware configuration, we choose 1 master and 1 core node of the instance type m5.xlarge
    e. Enable auto scaling option
    f. Choose your keypair and leave permissions to default.

4. Choose Create cluster.





- Cluster creation tasks takes 5-10 minutes . You may need to choose the refresh icon on the right or refresh your browser to receive updates.
- Under Network and hardware, find the Master and Core instance status. The status goes from Provisioning to Bootstrapping to Waiting during the cluster creation process. For more information, see Understanding the Cluster Lifecycle.
- As soon as you see the links for Security groups for Master and Security Groups for Core & Task, you can move on to the next step, but you may want to wait until the cluster starts successfully and is in the Waiting state.

Clone | Terminate | AWS CLI export

Cluster: dezyre-cluster Starting

Summary | Application user interfaces | Monitoring | Hardware | Configurations | Events | Steps | Bootstrap actions

Summary

ID: j-36W2U13LMG7WU
Creation date: 2020-10-31 20:27 (UTC+5:30)
Elapsed time: 1 second
After last step completes: Cluster waits
Termination protection: Off Change
Tags: -- View All / Edit
Master public DNS: --

Configuration details

Release label: emr-5.31.0
Hadoop distribution: Amazon 2.10.0
Applications: Hive 2.3.7, Hue 4.7.1, Mahout 0.13.0, Pig 0.17.0, Tez 0.9.2
Log URI: s3://dezyre-bucket/
EMRFS consistent view: Disabled
Custom AMI ID: --

Application user interfaces

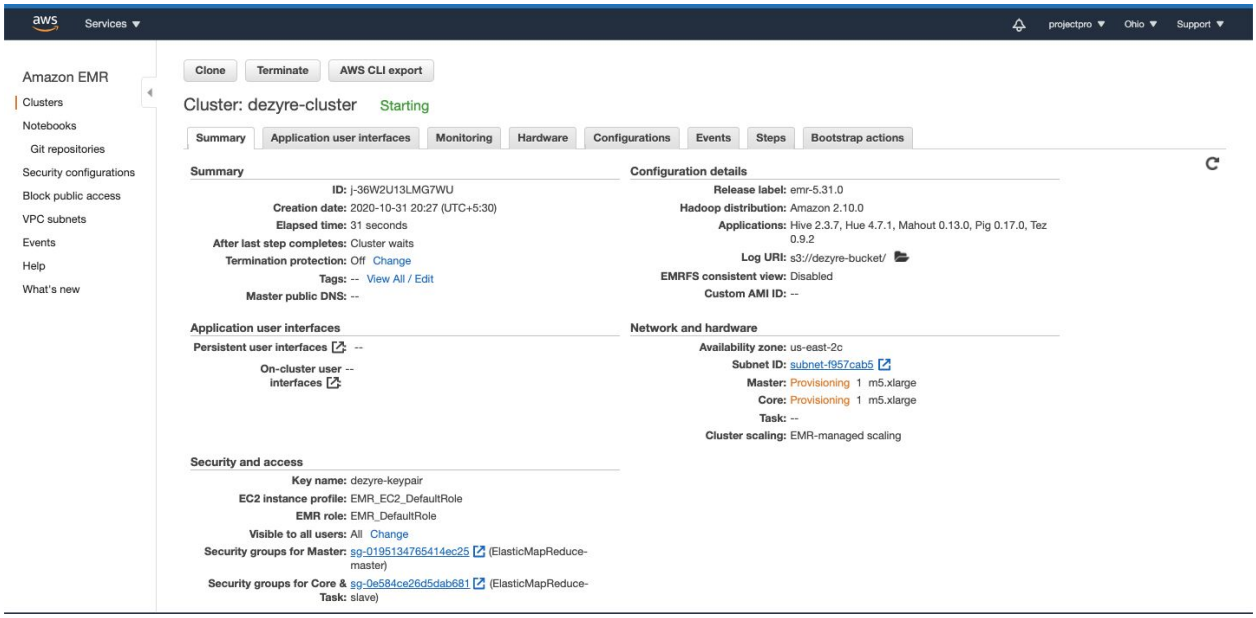
Persistent user interfaces --
On-cluster user interfaces --

Network and hardware

Availability zone: --
Subnet ID: subnet-f957cab5
Master: Provisioning 1 m5.xlarge
Core: Provisioning 1 m5.xlarge
Task: --
Cluster scaling: EMR-managed scaling

Security and access

Key name: dezyre-keypair
EC2 instance profile: EMR_EC2_DefaultRole
EMR role: EMR_DefaultRole
Visible to all users: All Change
Security groups for Master:
Security groups for Core & Task:

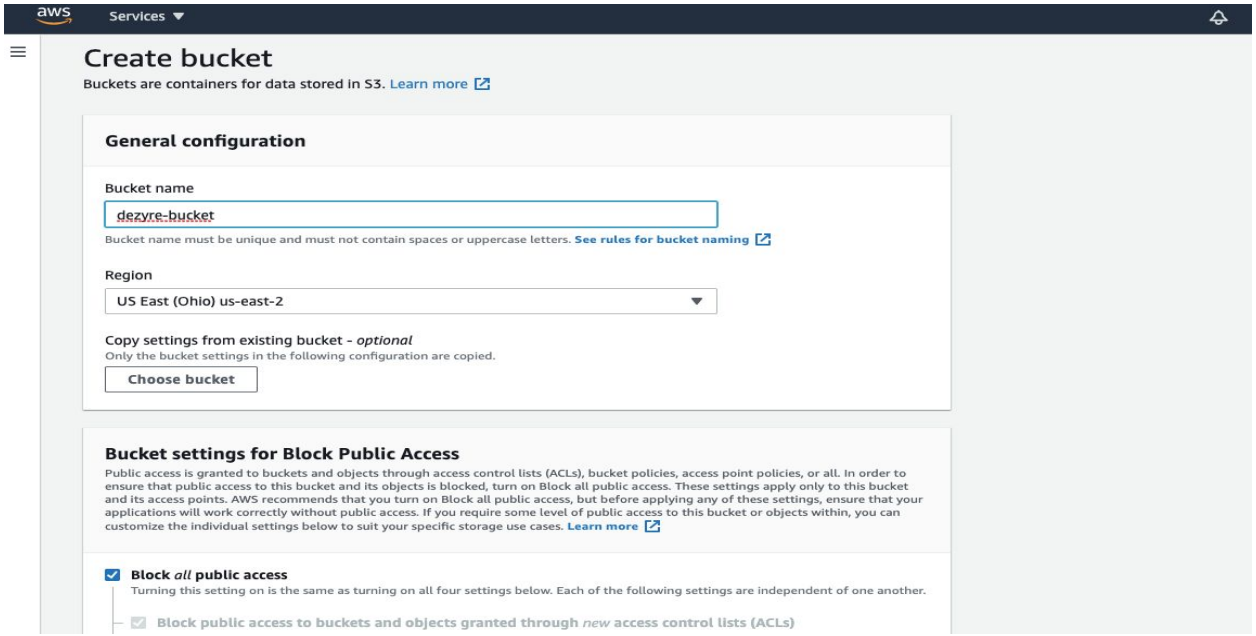


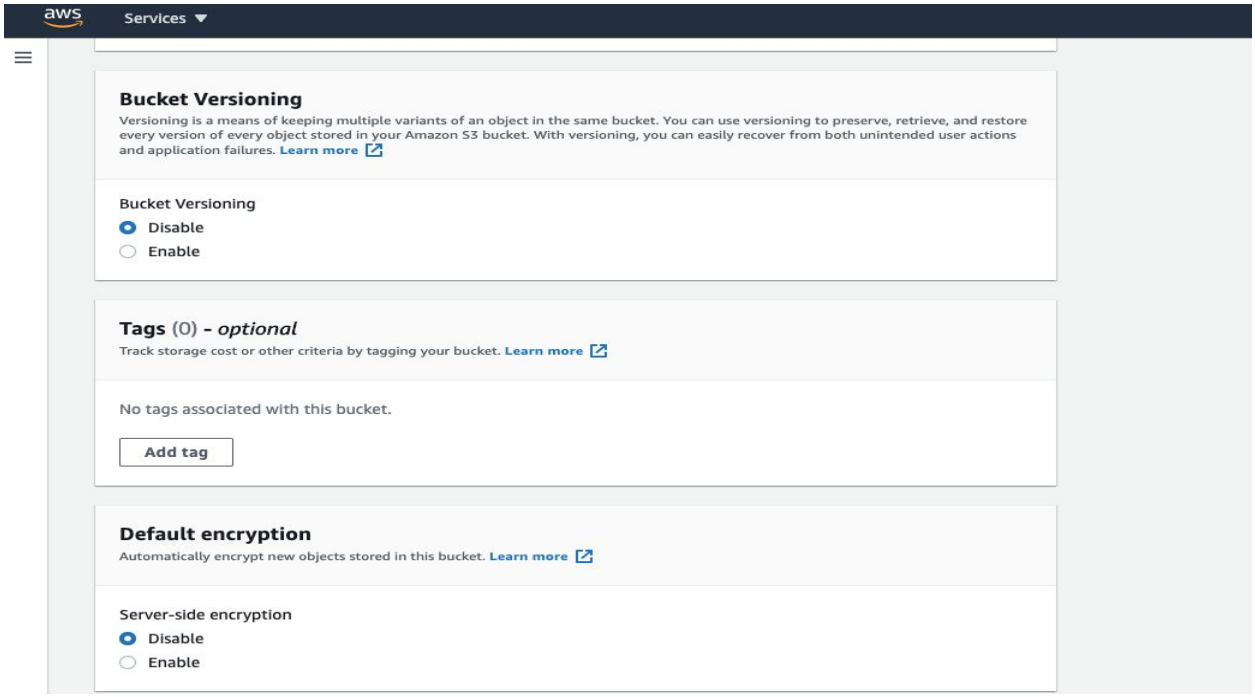
Step 2: Create S3 bucket

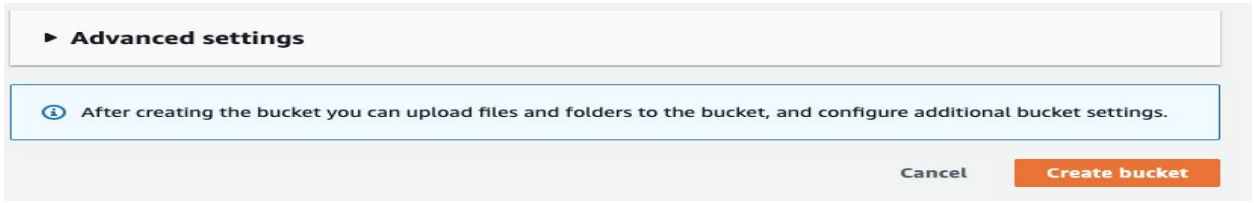
Give a unique name to the S3 (please add steps) to create and upload file

1. Sign in to the AWS Management Console and open the Amazon S3 console
2. Choose Create Bucket.
   The Create a Bucket dialog box opens.
3. Enter a bucket name, such as **dezyre-bucket.**
   This name should be globally unique, and cannot be the same name used by another bucket.
4. Select the Region for your bucket. To avoid paying cross-region bandwidth charges, create the Amazon S3 bucket in the same region as your cluster.
5. Choose Create.

You created a bucket with the URI s3n://dezyre-bucket/ as shown in the below.







When Amazon S3 successfully creates your bucket, the console displays your empty bucket in the Buckets panel.

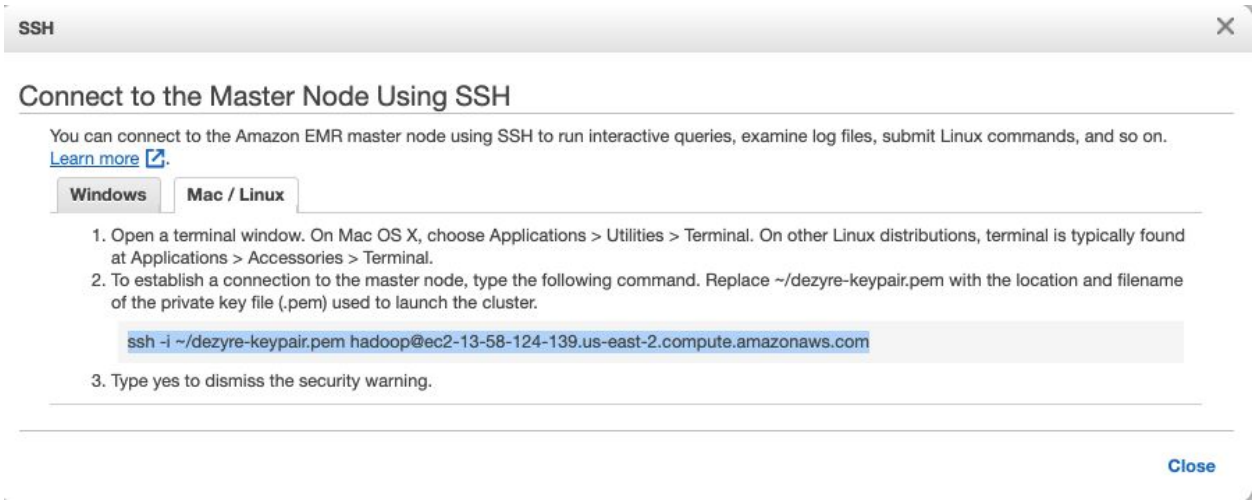
1. Create a folder.
    a. Click the name of the new bucket.
    b. Click the Actions button, and click **Create Folder** in the drop-down list.
    c. Name the new folder **load.**

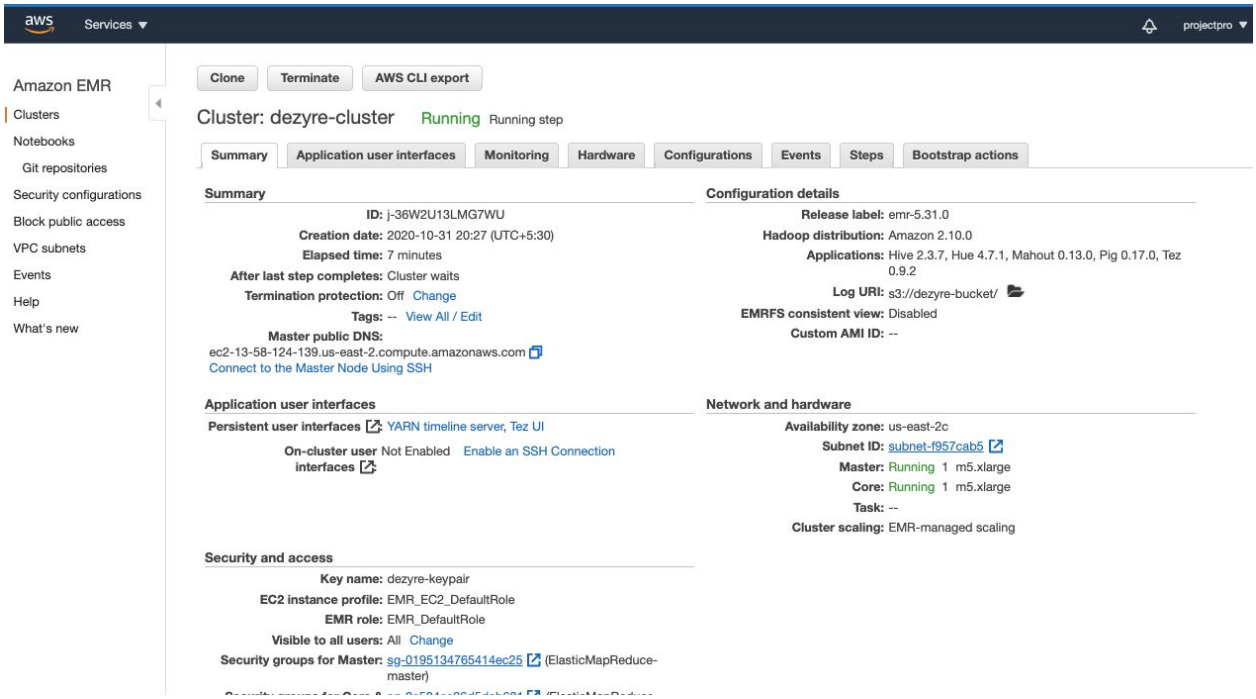
Upload the data files to the new Amazon S3 bucket.

a. Click the name of the data folder.
b. In the Upload - Select Files wizard, click **Add Files**.
   A file selection dialog box opens.
c. Select all of the files you downloaded and extracted, and then click **Open**.
d. Click **Start Upload**.

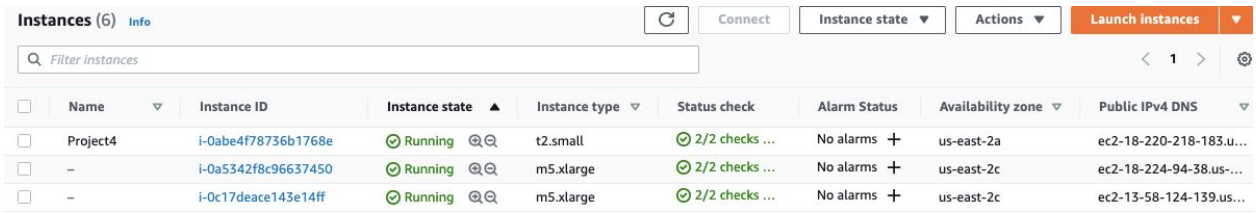
Here after creating the s3 will store data in a bucket.

Step 3: Connect to the emr master node using SSH

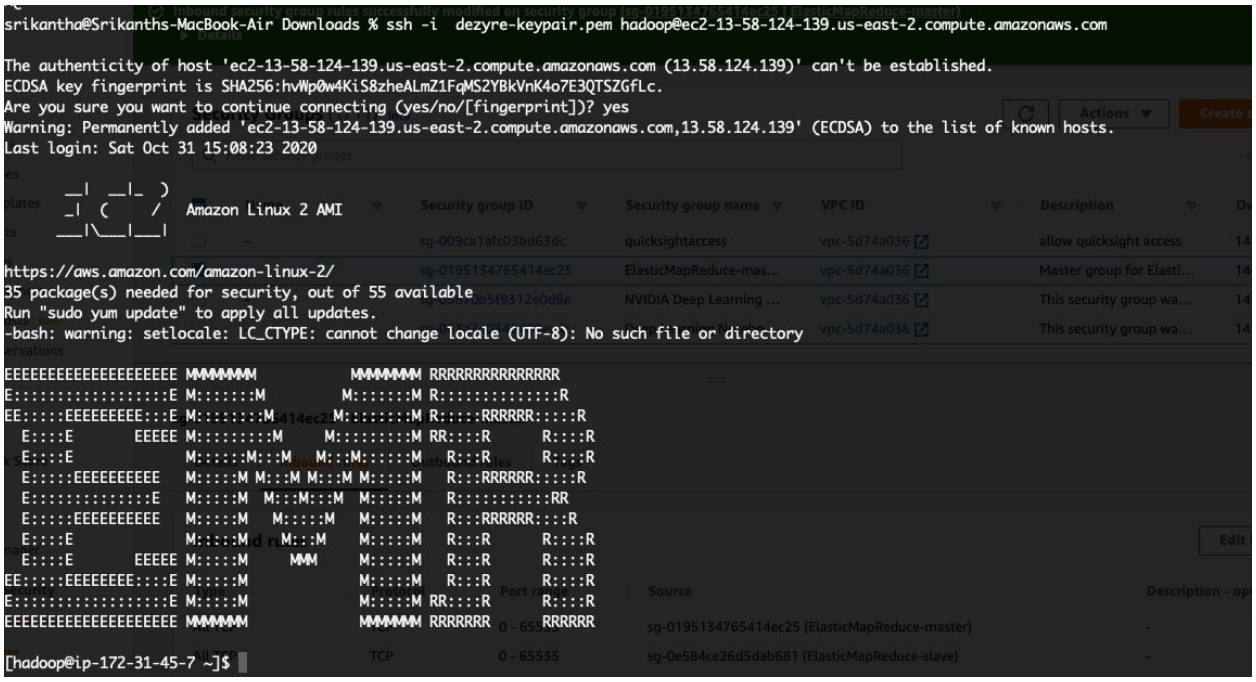




Once we open EC2 instance view, we would see the instances created and running



Now, ssh to emr master node as follows :



Step 4 : Creating the External Table from data in Amazon S3

Outside data sources are referenced in Amazon EMR by creating an EXTERNAL TABLE. This simply creates a reference to the data; no data is moved yet.

1. Once logged into the master node, start the Hive shell:

```
$ hive
```

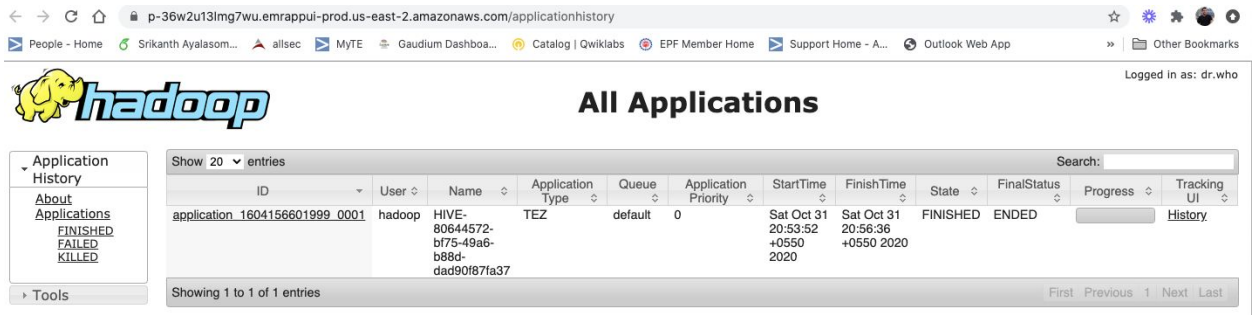
2. Define the source using a CREATE TABLE statement. For this example,

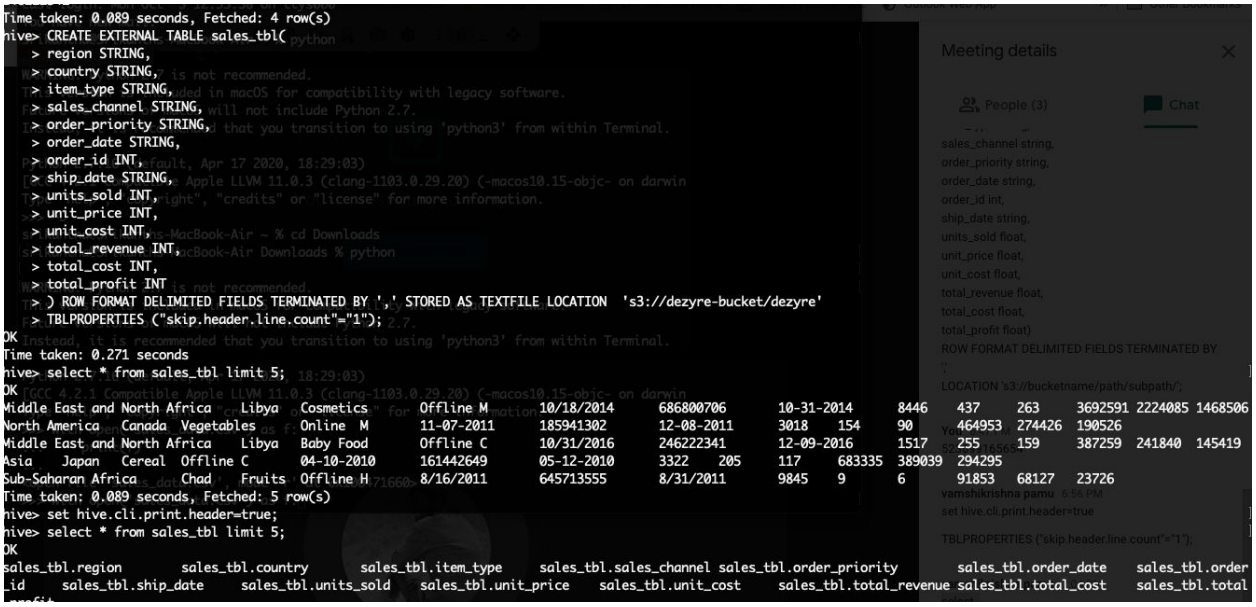
```
[hadoop@ip-172-31-45-7 ~]$ hive

Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: false
hive>
```

```
[hadoop@ip-172-31-45-7 ~]$ hive
Logging initialized using configuration in file:/etc/hive/conf.dist/hive-log4j2.properties Async: false
hive> show databases;
OK
default
Time taken: 0.523 seconds, Fetched: 1 row(s)
hive> use default;
OK
Time taken: 0.06 seconds
hive> show tables;
OK
Time taken: 0.065 seconds
hive>
```

```
hive> CREATE EXTERNAL TABLE IF NOT EXISTS sales_table(
    > region string,
    > country string,
    > item_type string,
    > sales_channel string,
    > order_priority string,
    > order_date string,
    > order_id STRING,
    > ship_date string,
    > units_sold STRING,
    > unit_price STRING,
    > unit_cost STRING,
    > total_revenue STRING,
    > total_cost STRING,
    > total_profit STRING
    > )
    > ROW FORMAT DELIMITED
    > FIELDS TERMINATED BY ','
    > LOCATION
    >  's3://dezyre-bucket/';
OK
Time taken: 2.69 seconds
hive>
```





Here we do using the external table processing the data parsing the columns and creating the new final table and storing data into it.

--we add scripts insert statement and final table here.

Step 5 : Using Tableau to Visualize the Data from EMR HIVE

For the next steps, you'll need the Tableau Desktop trial version installed on a Windows or MacOSX machine.
To connect the Tableau desktop tp the EMR Hadoop Hive on AWS, please follow the steps below:

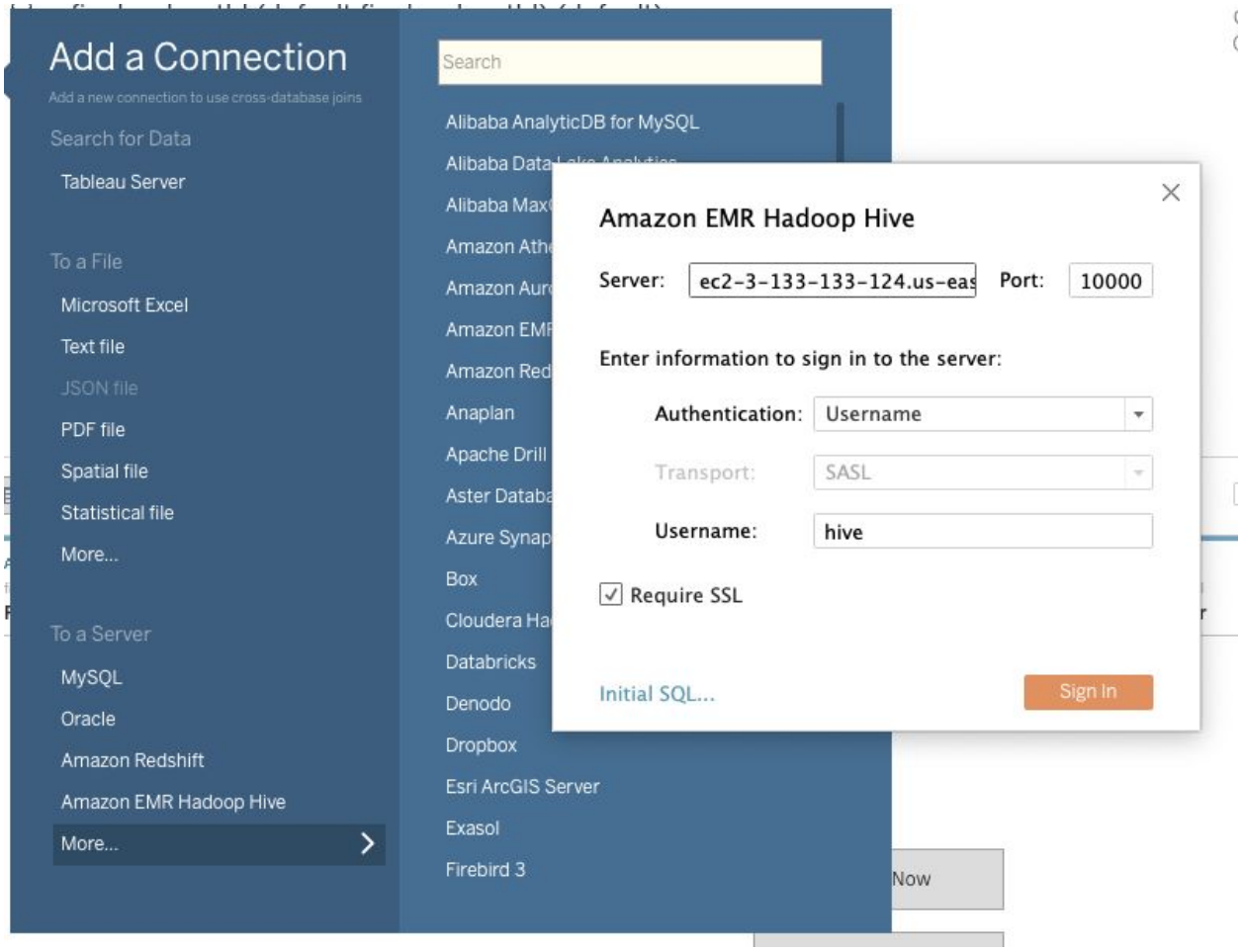
1. Install the ODBC driver on your machine with Tableau Desktop, required for connecting Tableau Desktop to HIVE on Amazon EMR.

a. Download the drivers.
b. Unzip the downloaded file. This should create a folder named "AmazonEMRHadoopHiveODBC."
c. Navigate to the required package for installing the driver.
Windows: AmazonEMRHadoopHive.msi
MacOSX:AmazonEMRHadoopHive.dmg
d. Run the package above and follow the prompts to install the ODBC driver.

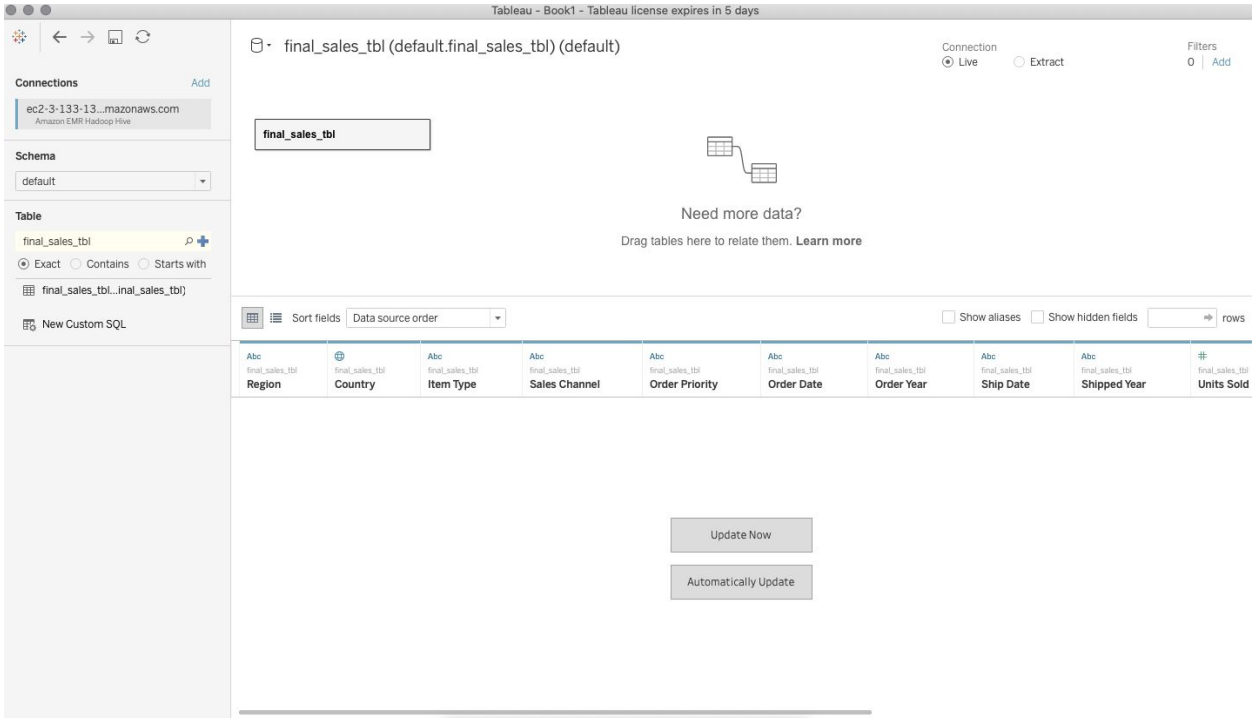
2. Modify the Amazon EMR cluster's Master Security Group so Tableau can connect with the AmazonEMRHadoopHive server running on the master node of the Amazon EMR cluster.

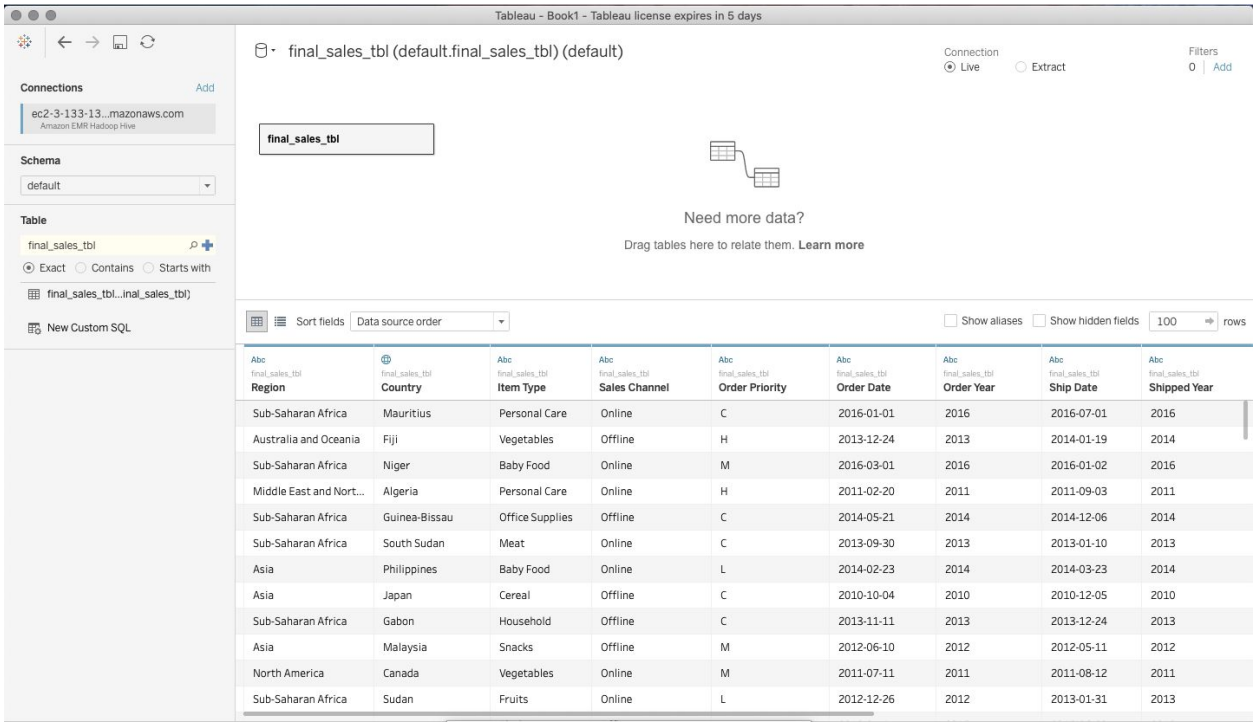
a. Click the Amazon EC2 tab in the AWS Management Console to open the Amazon EC2 console.
b. In the navigation pane, select Security Groups under the Network and Security group.
c. In the Security Groups list, select Elastic MapReduce-master.
d. In the lower pane, click the Inbound tab.
e. In the Port Range field type 10000. Leave the default value in the Source field.
f. Click Add Rule, and then click Apply Rule Changes.

3. Follow the steps as directed by Tableau to enable Amazon EMR Hadoop Hive as a data connection option in Tableau.
4. Got to Add a connextion -> Click on more-> Select Amazon EMR Hadoop Hive and a pop up window appears to enter details .
5. Give your server DNS , port number as 10000
6. Choose Authentication as Username from the dropdown
7. Specify Username as hive
8. Check the RequireSSL box and click on Sign in



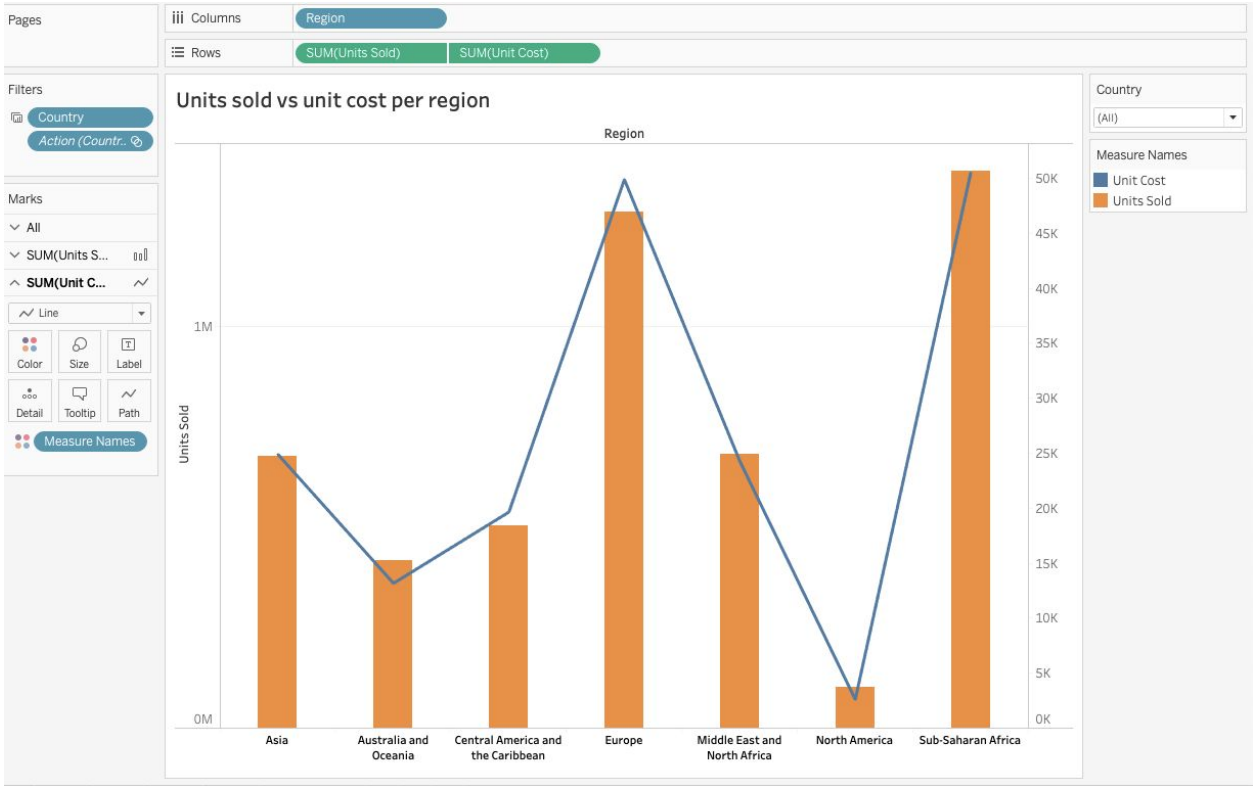
- Search for your Database and press enter
- Then search for your schema and press enter
- Drag the table in to work area in workbook in tableau as shown in the below images
- Now, navigate to your sheets and make different visualisations

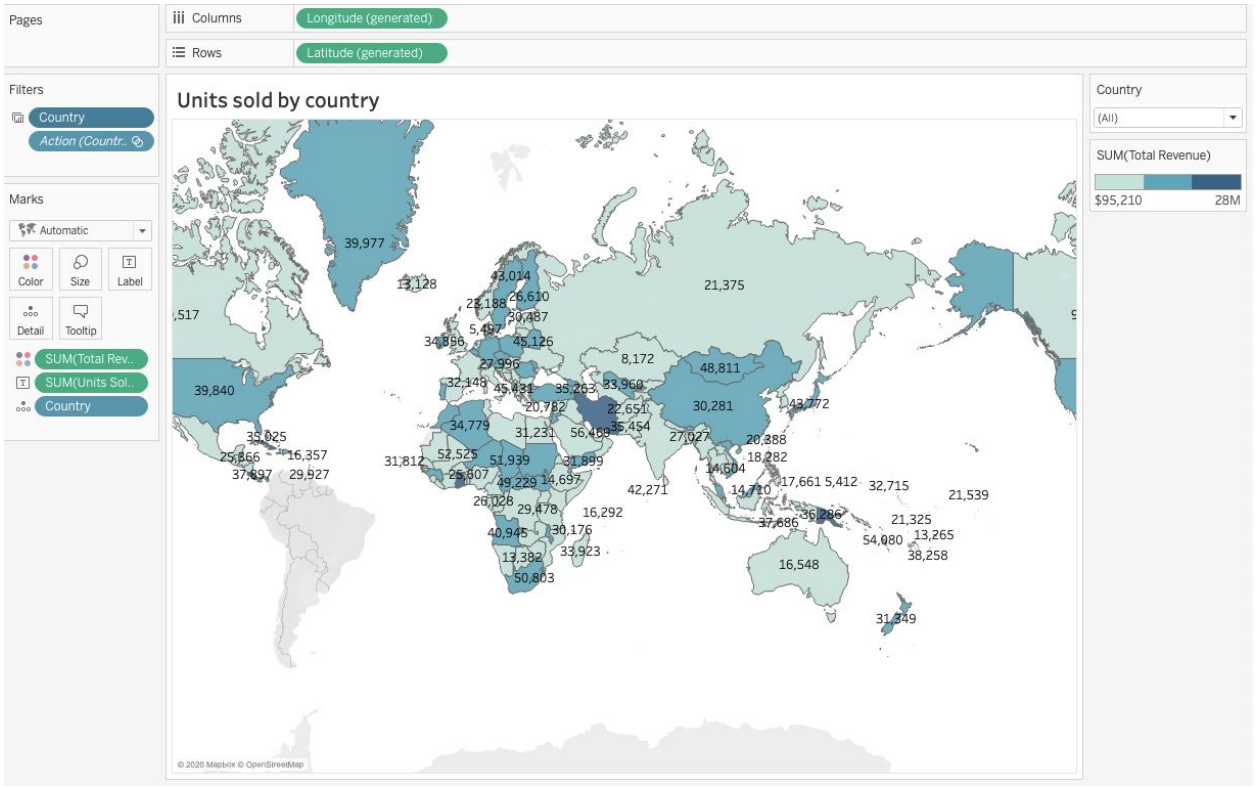


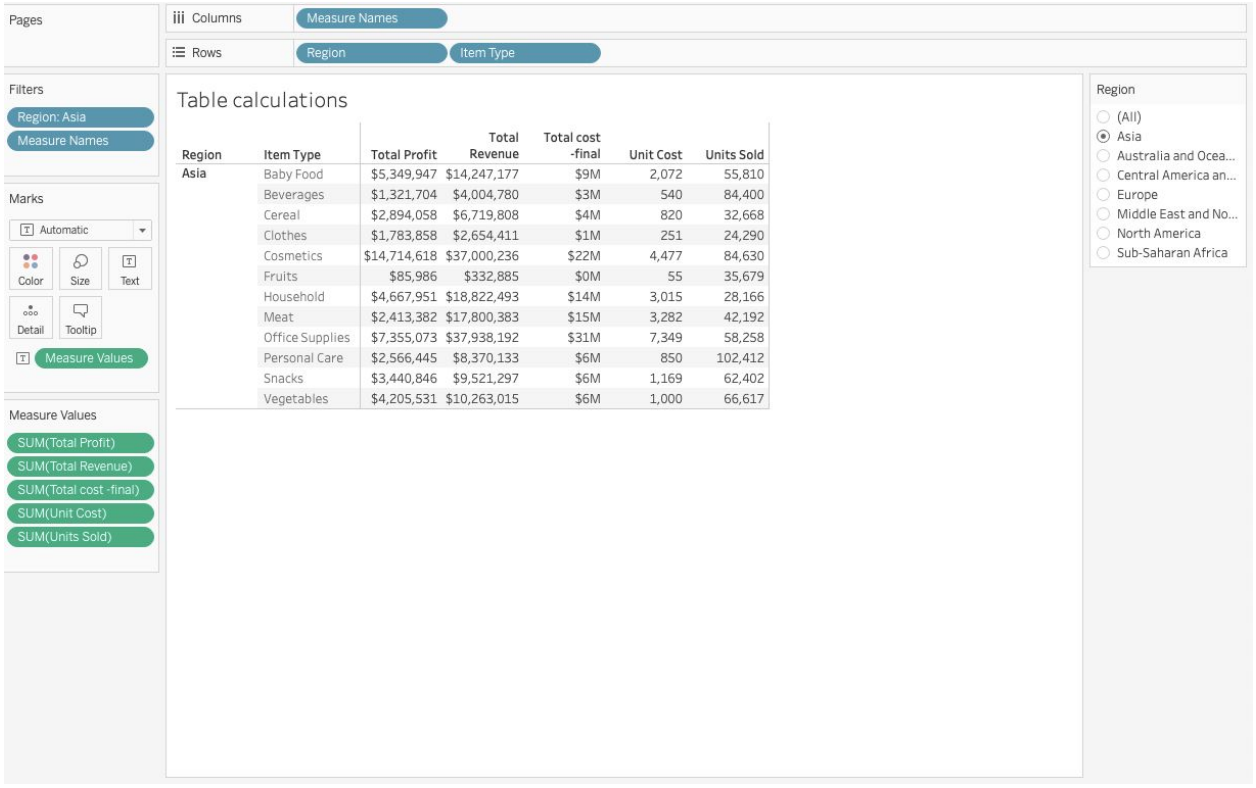


Region	Country	Item Type	Sales Channel	Order Priority	Order Date	Order Year	Ship Date	Shipped Year
Sub-Saharan Africa	Mauritius	Personal Care	Online	C	2016-01-01	2016	2016-07-01	2016
Australia and Oceania	Fiji	Vegetables	Offline	H	2013-12-24	2013	2014-01-19	2014
Sub-Saharan Africa	Niger	Baby Food	Online	M	2016-03-01	2016	2016-01-02	2016
Middle East and Nort...	Algeria	Personal Care	Online	H	2011-02-20	2011	2011-09-03	2011
Sub-Saharan Africa	Guinea-Bissau	Office Supplies	Offline	C	2014-05-21	2014	2014-12-06	2014
Sub-Saharan Africa	South Sudan	Meat	Online	C	2013-09-30	2013	2013-01-10	2013
Asia	Philippines	Baby Food	Online	L	2014-02-23	2014	2014-03-23	2014
Asia	Japan	Cereal	Offline	C	2010-10-04	2010	2010-12-05	2010
Sub-Saharan Africa	Gabon	Household	Offline	C	2013-11-11	2013	2013-12-24	2013
Asia	Malaysia	Snacks	Offline	M	2012-06-10	2012	2012-05-11	2012
North America	Canada	Vegetables	Online	M	2011-07-11	2011	2011-08-12	2011
Sub-Saharan Africa	Sudan	Fruits	Online	L	2012-12-26	2012	2013-01-31	2013

Creating different types of visualizations on sales data using tableau

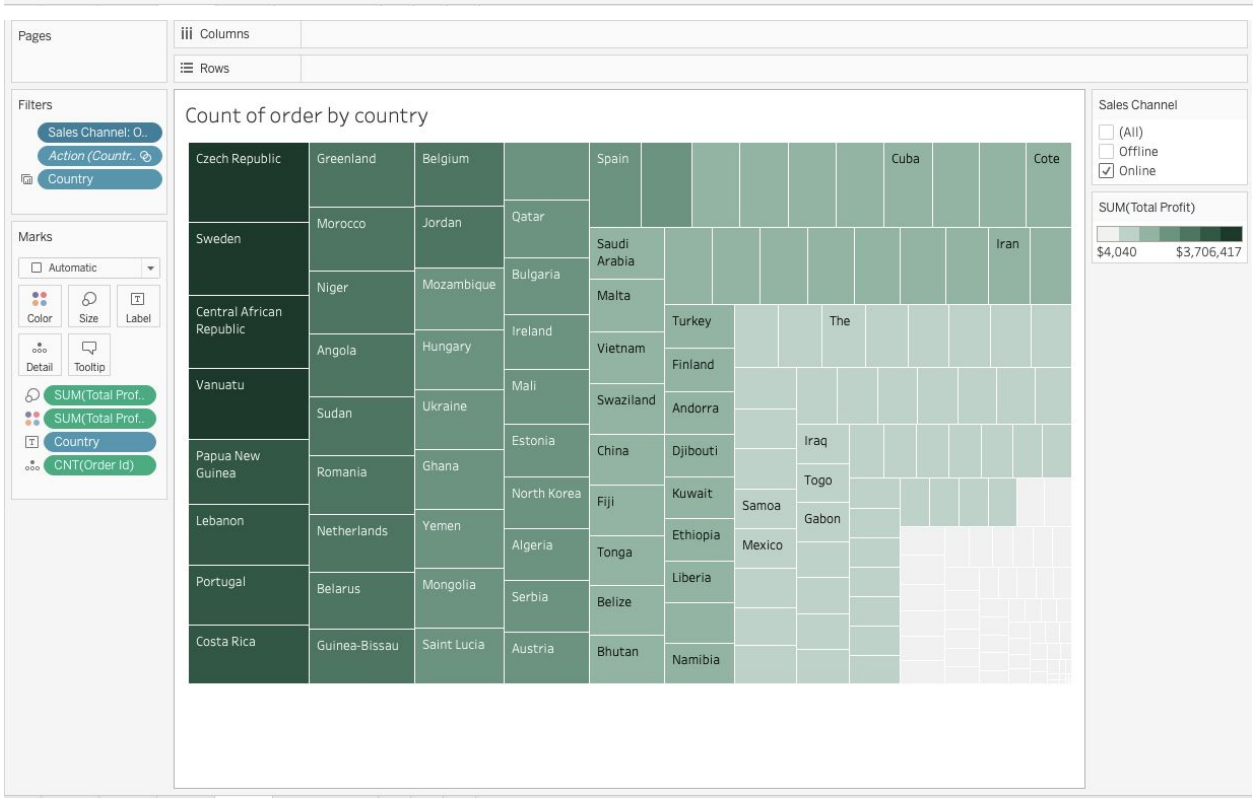






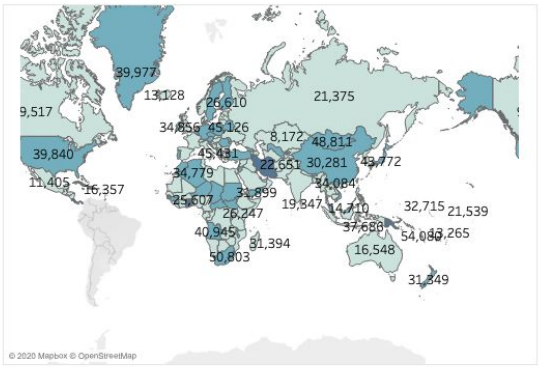
Region	Item Type	Total Profit	Total Revenue	Total cost -final	Unit Cost	Units Sold
Asia	Baby Food	$5,349,947	$14,247,177	$9M	2,072	55,810
	Beverages	$1,321,704	$4,004,780	$3M	540	84,400
	Cereal	$2,894,058	$6,719,808	$4M	820	32,668
	Clothes	$1,783,858	$2,654,411	$1M	251	24,290
	Cosmetics	$14,714,618	$37,000,236	$22M	4,477	84,630
	Fruits	$85,986	$332,885	$0M	55	35,679
	Household	$4,667,951	$18,822,493	$14M	3,015	28,166
	Meat	$2,413,382	$17,800,383	$15M	3,282	42,192
	Office Supplies	$7,355,073	$37,938,192	$31M	7,349	58,258
	Personal Care	$2,566,445	$8,370,133	$6M	850	102,412
	Snacks	$3,440,846	$9,521,297	$6M	1,169	62,402
	Vegetables	$4,205,531	$10,263,015	$6M	1,000	66,617


Pages
Columns
QUARTER(Order D..
Rows
SUM(Total Profit)
SUM(Total Revenue)
Filters
Sales Channel: Offline
Region: Australia and..
Marks
All
SUM(Total Pr...
SUM(Total R...
Automatic
Color
Size
Label
Detail
Tooltip
Path
Measure Names
Revenue vs profit By region and sales channel
$8.0M
$7.0M
$6.0M
$5.0M
$4.0M
$3.0M
$2.0M
$1.0M
$0.0M
$2.0M
$1.5M
$1.0M
$0.5M
$0.0M
Total Profit
Total Revenue
2010 Q3
2011 Q3
2012 Q3
2013 Q3
2014 Q3
2015 Q3
2016 Q3
2017 Q3
Quarter of Order Date
Region
(All)
Asia
Australia and Ocea...
Central America an...
Europe
Middle East and No...
North America
Sub-Saharan Africa
Sales Channel
(All)
Offline
Online
Measure Names
Total Profit
Total Revenue


Pages
Columns
Rows
Filters
Sales Channel: O..
Action (Countr..
Country
Marks
Automatic
Color
Size
Label
Detail
Tooltip
SUM(Total Prof..
SUM(Total Prof..
Country
CNT(Order Id)
Count of order by country
Czech Republic
Sweden
Central African Republic
Vanuatu
Papua New Guinea
Lebanon
Portugal
Costa Rica
Greenland
Morocco
Niger
Angola
Sudan
Romania
Netherlands
Belarus
Guinea-Bissau
Belgium
Jordan
Mozambique
Hungary
Ukraine
Ghana
Yemen
Mongolia
Saint Lucia
Qatar
Bulgaria
Ireland
Mali
Estonia
North Korea
Algeria
Serbia
Austria
Spain
Saudi Arabia
Malta
Vietnam
Swaziland
China
Fiji
Tonga
Belize
Bhutan
Turkey
Finland
Andorra
Djibouti
Kuwait
Ethiopia
Liberia
Namibia
Samoa
Mexico
Iraq
Togo
Gabon
The
Cuba
Iran
Cote
Sales Channel
(All)
Offline
Online
SUM(Total Profit)
$4,040
$3,706,417

Sales dashboard

Units sold by country



Total revenue and cost per country

Filters

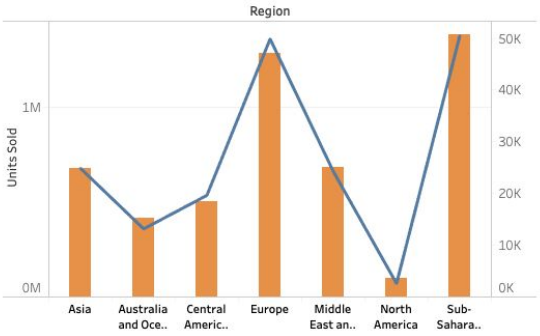
Sales Channel

- Offline
- Online

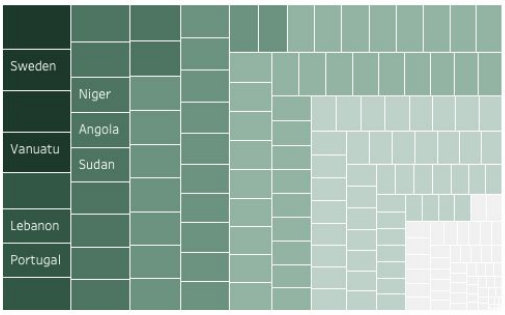
Country

(All)

Units sold vs unit cost per region



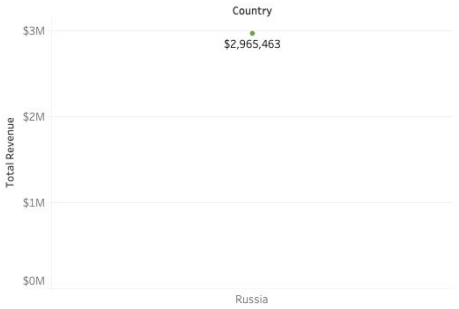
Count of order by country



Sales dashboard

Units sold by country

Total revenue and cost per country

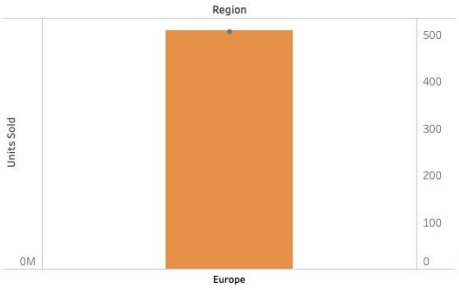


Filters

Sales Channel
- Offline
- Online

Country
(All)

Units sold vs unit cost per region



Count of order by country

Sales dashboard

Units sold by country

Total revenue and cost per country

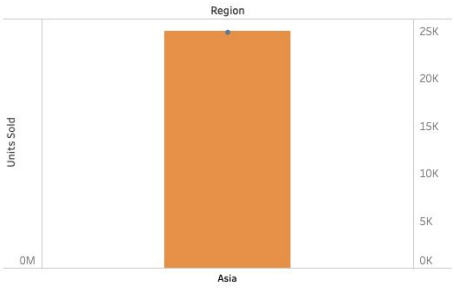
Filters

Sales Channel
- Offline
- Online

Country
(All)

Region
Asia

Units sold vs unit cost per region



Count of order by country