**Caterpillar INC-Team 6**



**Team Members:**

Purvang Jayesh Thakkar 001387983

Ira Pantbalekundri 001423854

Ashu Kapil 001400324

Caterpillar Inc. (Caterpillar) is a manufacturer of construction and mining equipment, diesel and natural gas engines, industrial gas turbines and diesel-electric locomotives.

We are focusing on one brand of the Caterpillar Inc which is CAT.

CAT is their main brand and its line of machines range from tracked tractors to [hydraulic excavators](https://en.wikipedia.org/wiki/Excavator), [backhoe loaders](https://en.wikipedia.org/wiki/Backhoe), [motor graders](https://en.wikipedia.org/wiki/Grader), off-highway trucks, [wheel loaders](https://en.wikipedia.org/wiki/Loader_(equipment)), agricultural [tractors](https://en.wikipedia.org/wiki/Tractor) and locomotives. Caterpillar machinery is used in the construction, [road-building](https://en.wikipedia.org/wiki/Road_construction), mining, [forestry](https://en.wikipedia.org/wiki/Forestry), energy, transportation and material-handling industries.

**Marketing Analytics System Architecture:**

We have incorporated Amazon Web Services (AWS) as the base system to perform marketing analytics.

Image-1

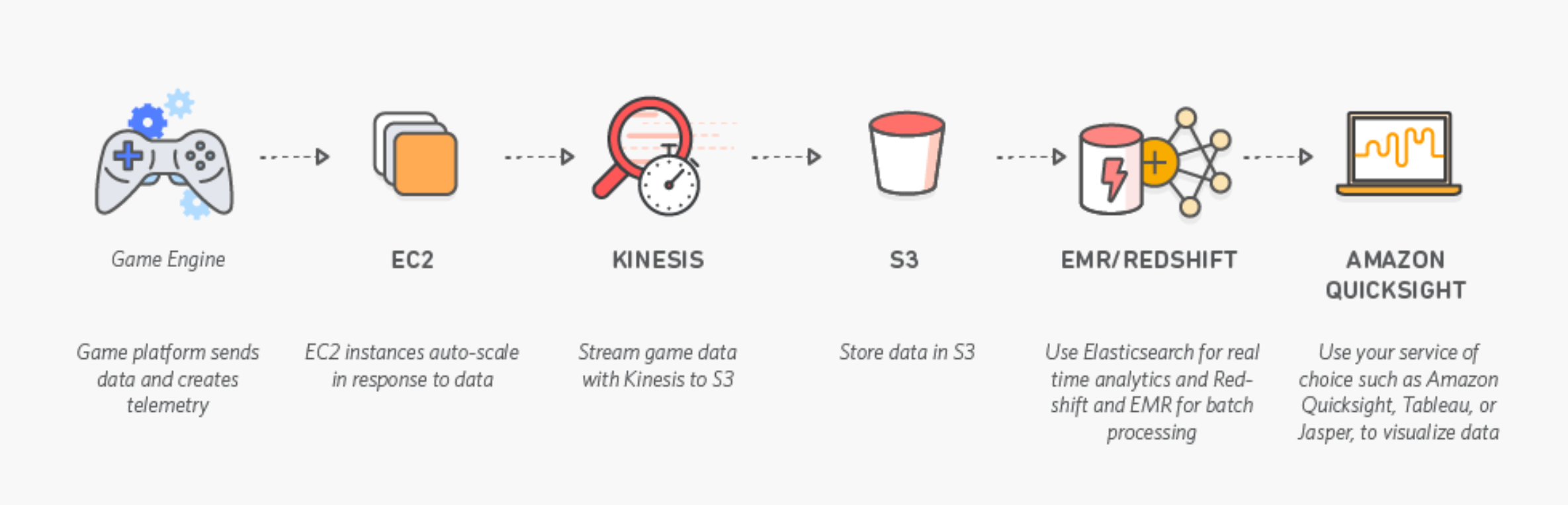
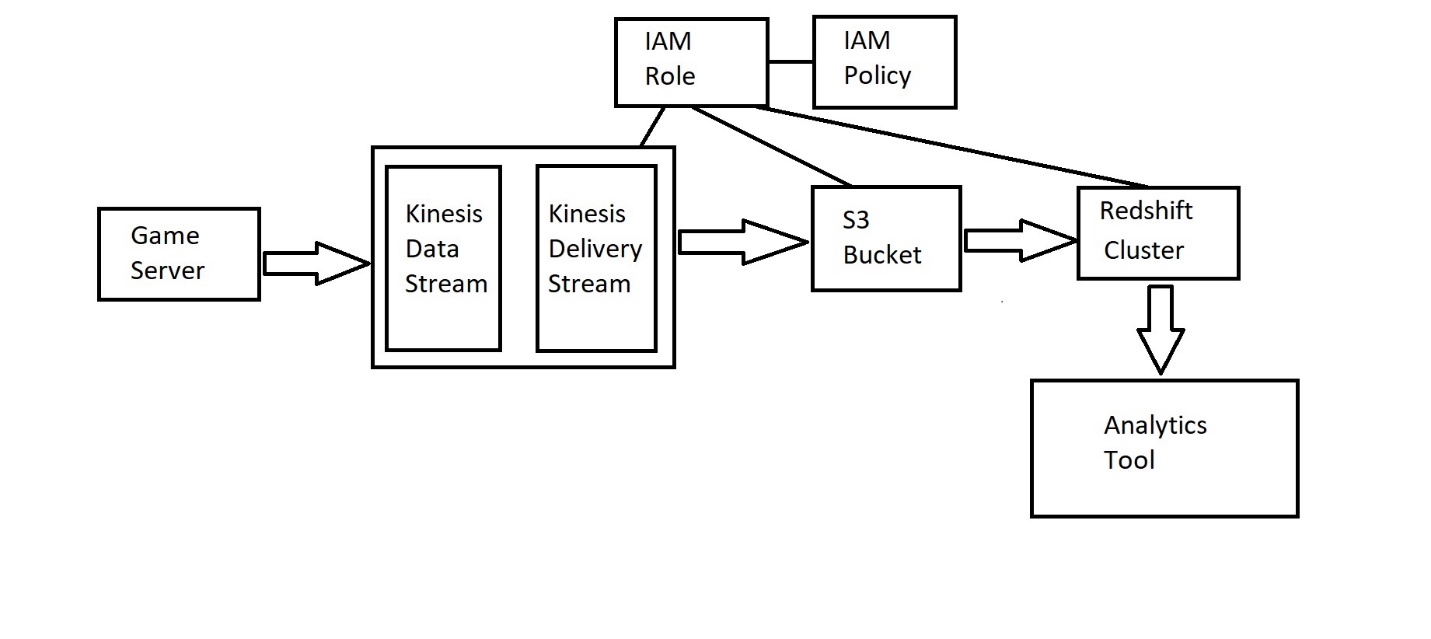


Image-2



The above 2 images depict the AWS architecture we will be implementing:

The first image is the default AWS gaming architecture pipeline template which we have later modified in the second image based on what we needed to implement the marketing analytics system.

The AWS CloudFormation template deploys an Amazon Virtual Private Cloud (Amazon VPC) network topology with two private subnets; two Amazon Kinesis streams; two KCL consumer applications running on top of AWS Elastic Beanstalk; an application for database tasks and maintenance; an Amazon Redshift cluster; and Amazon Simple Storage Service (Amazon S3) buckets for telemetry, errors, code, and configuration information; and an Amazon S3 endpoint.

Real time or batch time data of the company will be sent as an input in JSON-formatted/ csv file type to an Amazon Kinesis stream (called the *telemetry stream*) that collects and processes those events. A KCL consumer application (called the S3Connector) validates, sanitizes, and enriches the events, and archives the events as a batch telemetry file in Amazon S3. The application also sends a pointer to the location of the batch telemetry file to a separate Amazon Kinesis stream (called the *file stream*) that initiates the process of loading the data into Amazon Redshift.

Once the data has been loaded into Amazon Redshift, we will be using a business intelligence tool called **Tableau** to perform further analysis and **python** to gain more insights to the data which we have stored in the AWS redshift database.

We will be mimicking the Game server with the mock dataset of the Caterpillar INC company which comprises of the sales analysis data over the past 3 years.

Step1: Created an S3 Bucket in AWS which is used to Store, protect, and retrieve any amount of data, from anywhere.



Step2: Create IAM role. We needed to create 2 IAM roles. One for S3 bucket and one for Kinesis. An IAM role is an IAM identity that you can create in your account that has specific permissions. An IAM role is like an IAM user, in that it is an AWS identity with permission policies that determine what the identity can and cannot do in AWS. However, instead of being uniquely associated with one person, a role is intended to be assumable by anyone who needs it.

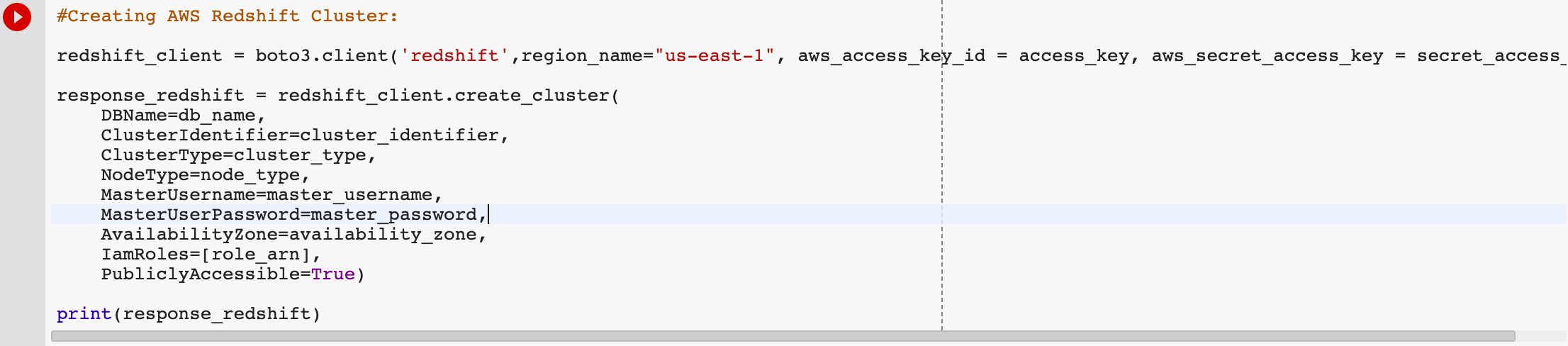


Step3: Creating IAM Policy to attach to the IAM roles created.

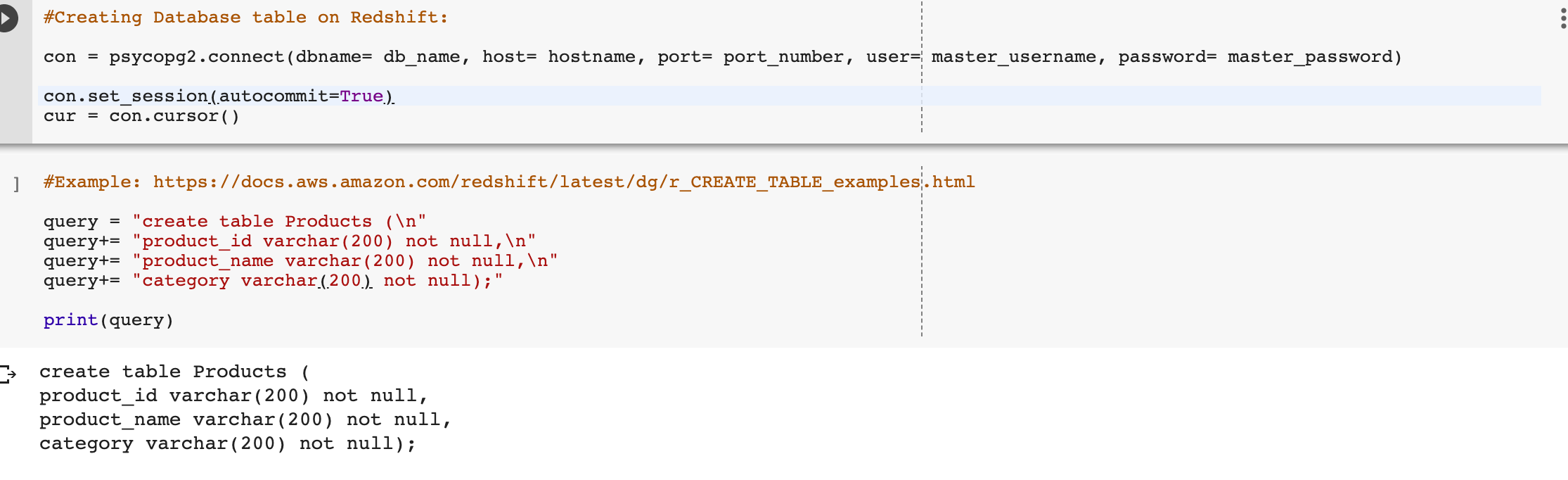


Step4: Creating an AWS Redshift Cluster

An Amazon Redshift data warehouse is a collection of computing resources called nodes, which are organized into a group called a cluster. Each cluster runs an Amazon Redshift engine and contains one or more databases.



Step5:Creating database tables in the Redshift Cluster



Data can be processed in 2 ways: Real time or Batch time. We have implemented both the ways and implemented the same.

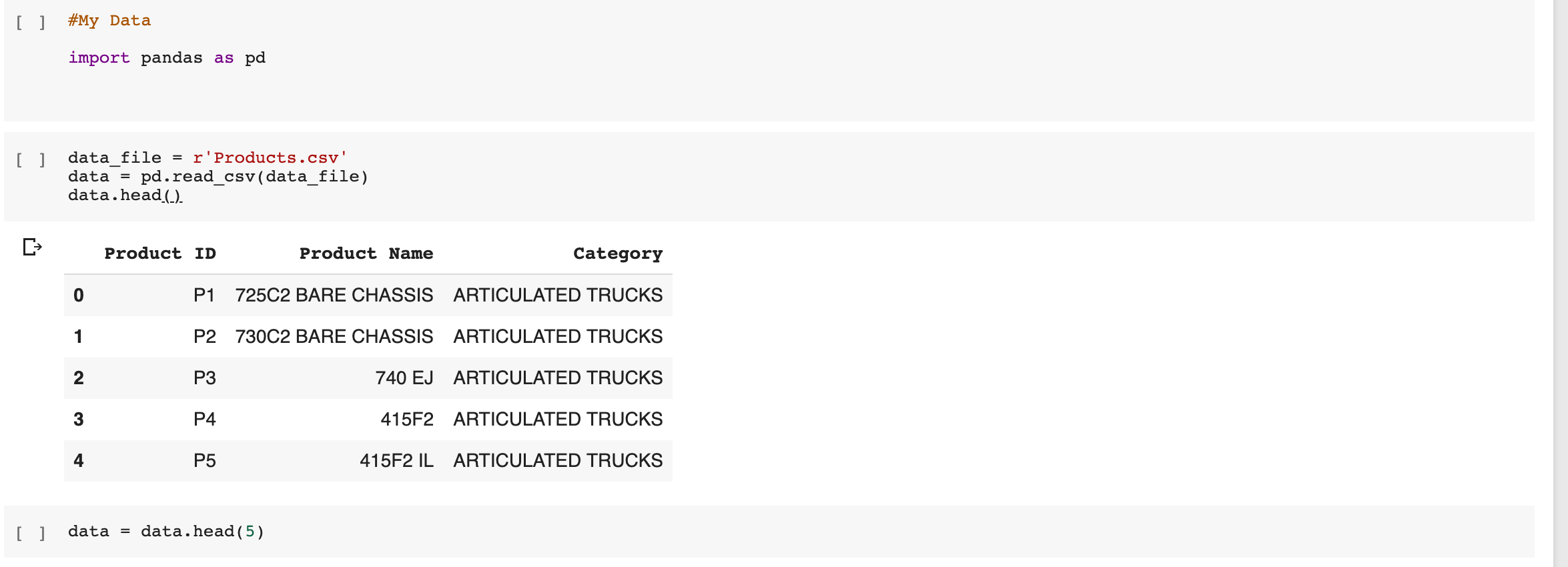
Step6:Creating kinesis stream for real time data flowing.



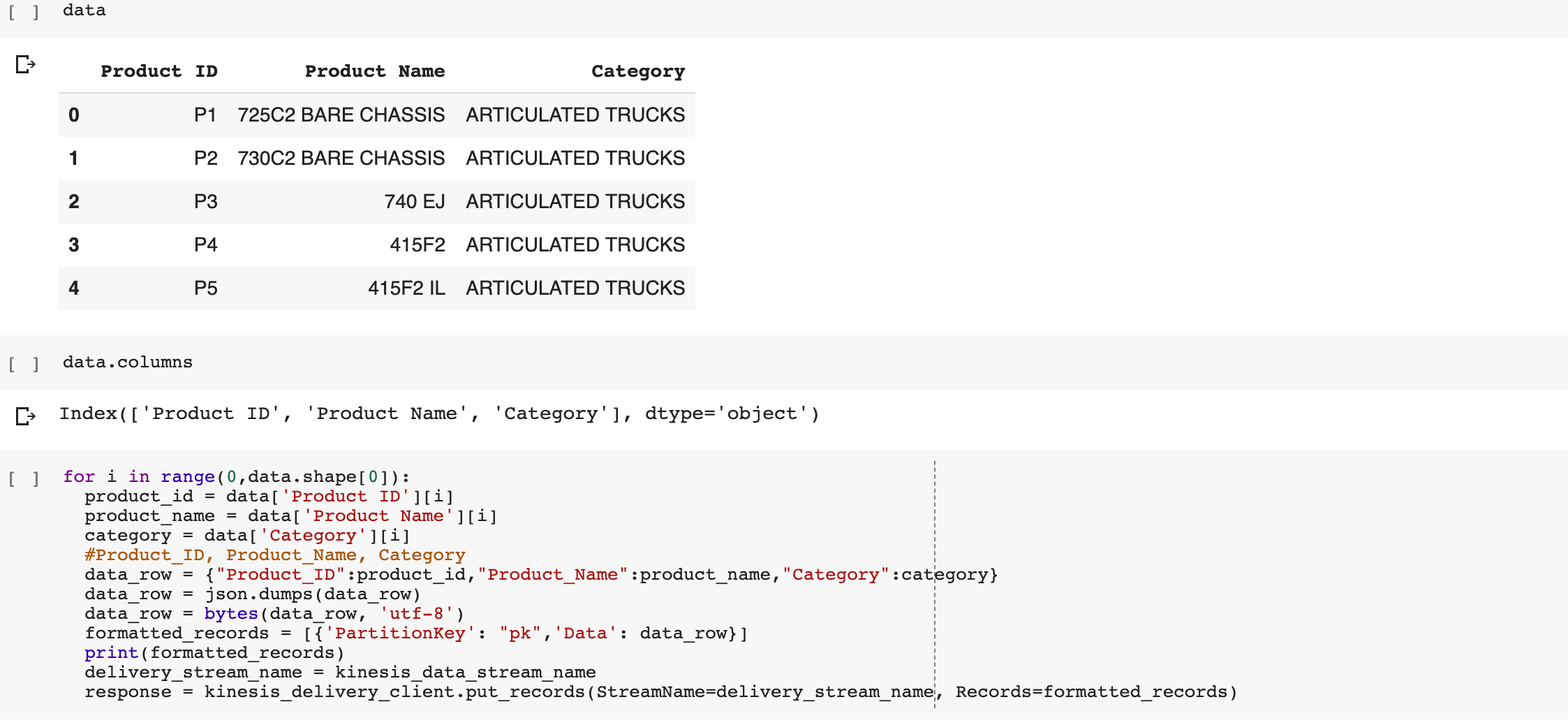
Amazon Kinesis makes it easy to collect, process, and analyze real-time, streaming data so you can get timely insights and react quickly to new information

Step7: Creating kinesis delivery stream.





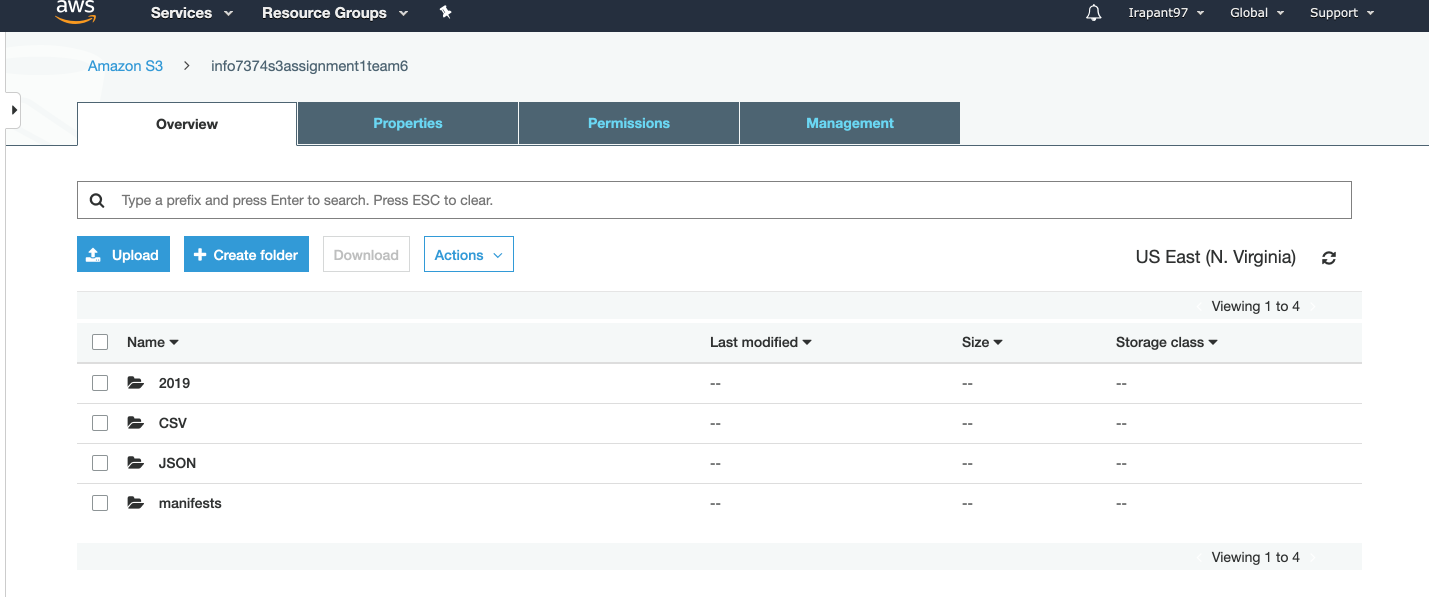


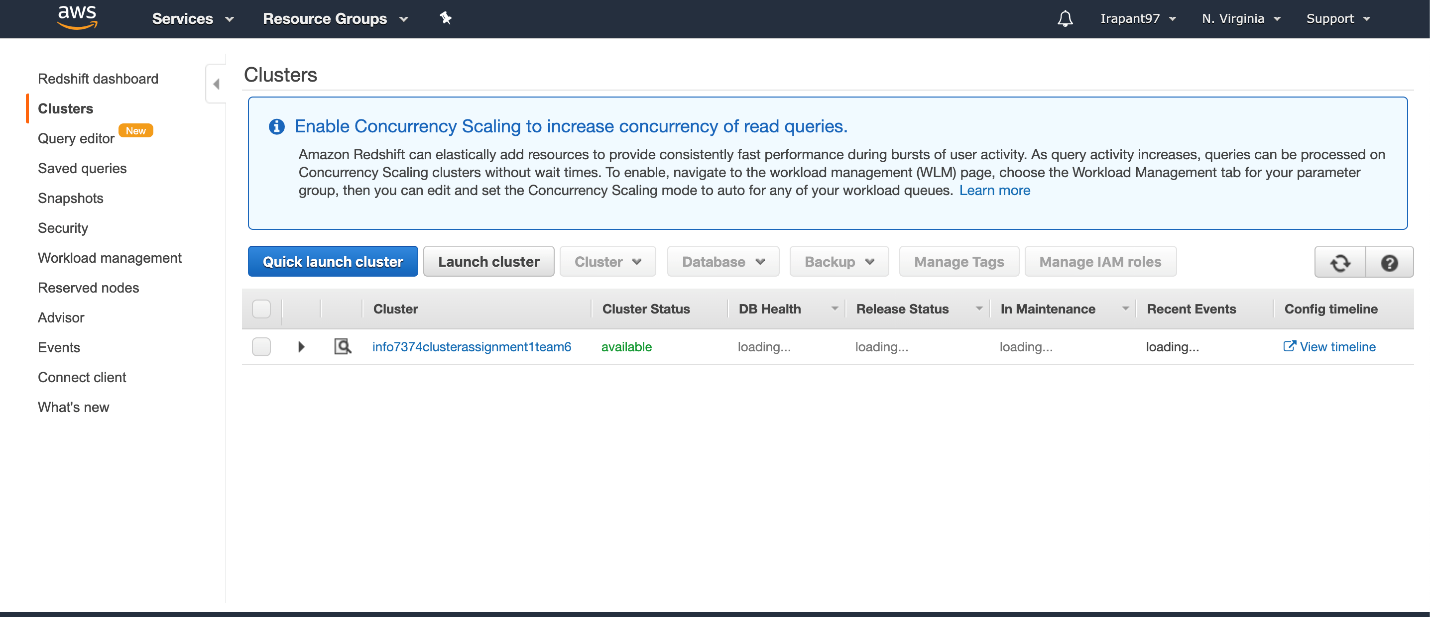


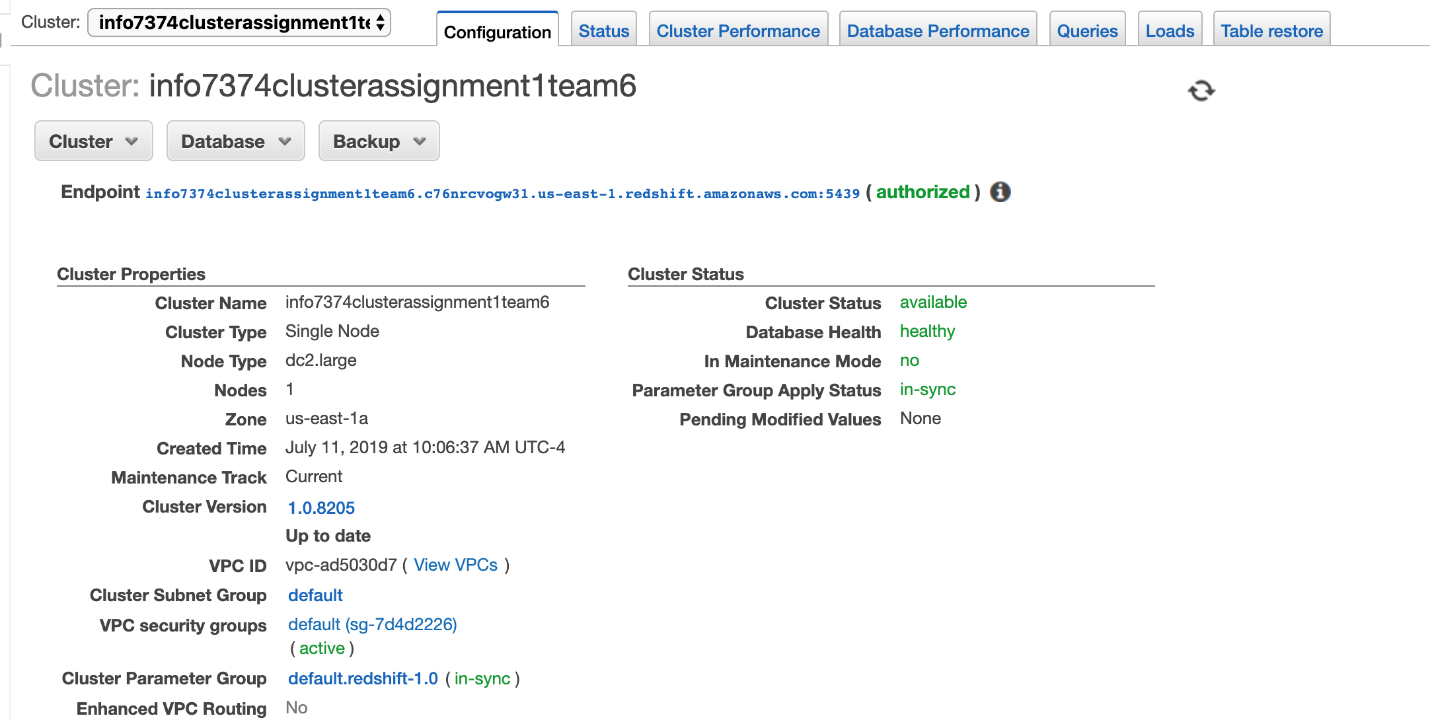


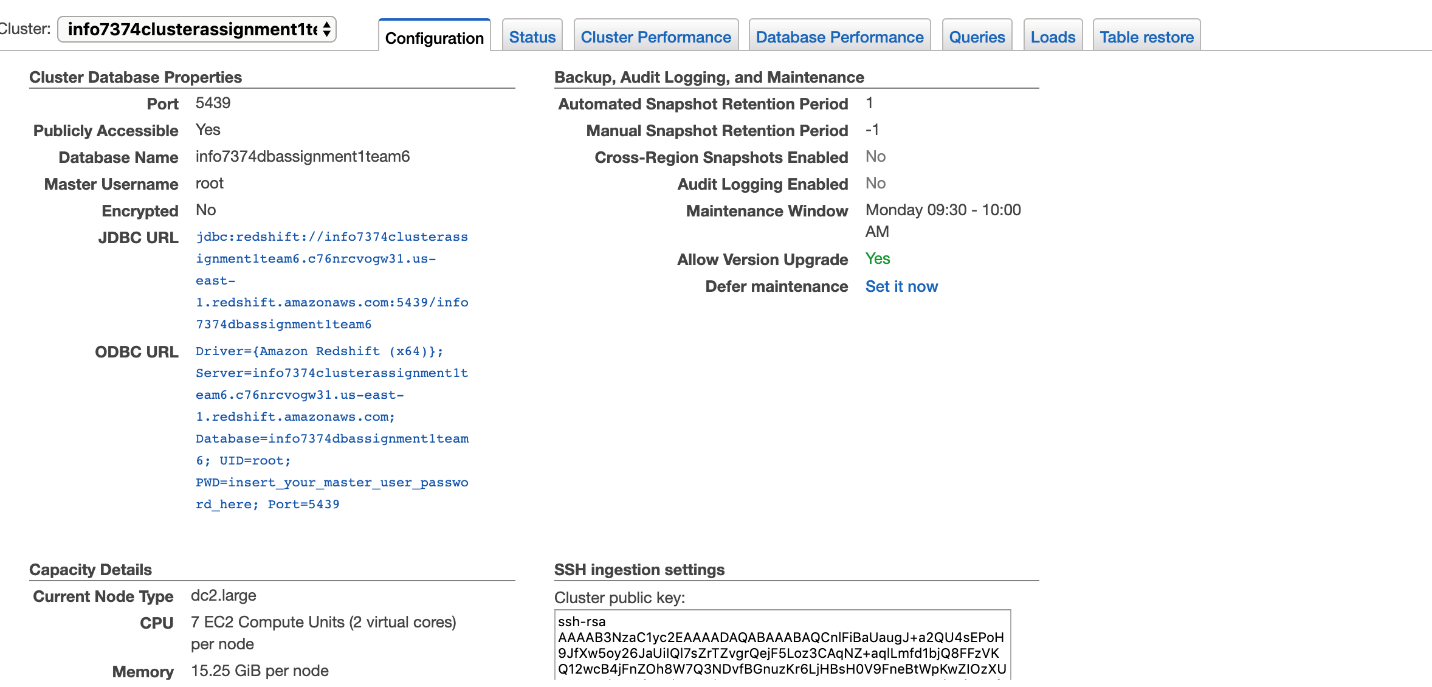
For Batch time processing of the data,

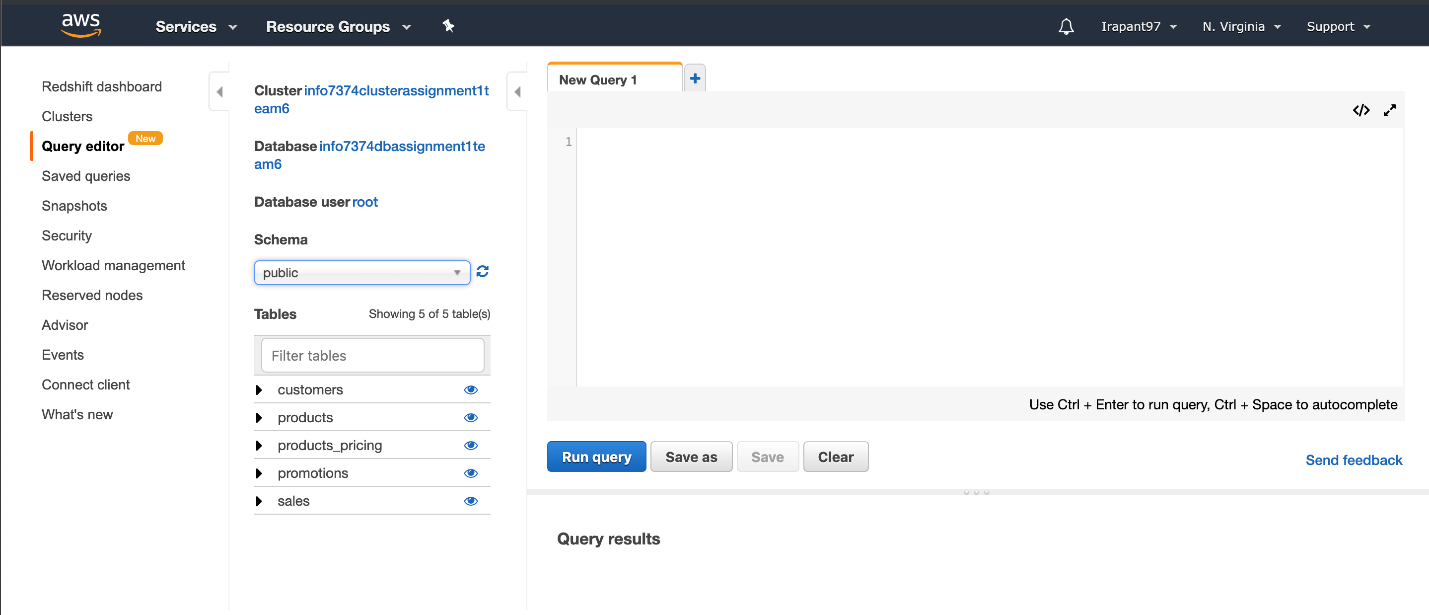
We would upload batches of file (in our scenario CSVs) in to S3 bucket. Once done we would populate it to the Redshift Table using the copy command.











**Challenge:**

Caterpillar Inc is currently facing majorly real-world issues with respect to the growth of the company.

First being that, since it’s range of products are mostly premium categorized and expensive they have their majority of customers from large scale industries and that’s where the most revenue comes from

Second being inconsistent promotional strategies of Caterpillar Inc which is hampering the sales of certain product categories.

So, we are going to come up with a detailed analysis and figure out how diagnose and find solutions to these problems Caterpillar Inc is facing using marketing Analytics.

**Dataset:**

Since we did not find any concrete dataset or any similar kind of dataset with the challenge our company faces, we created a mock dataset of our own.

The mock dataset comprises of Sales data of Caterpillar Inc for the past 3 years. It consists of the orders received, the products, customers data, promotions and product pricing.

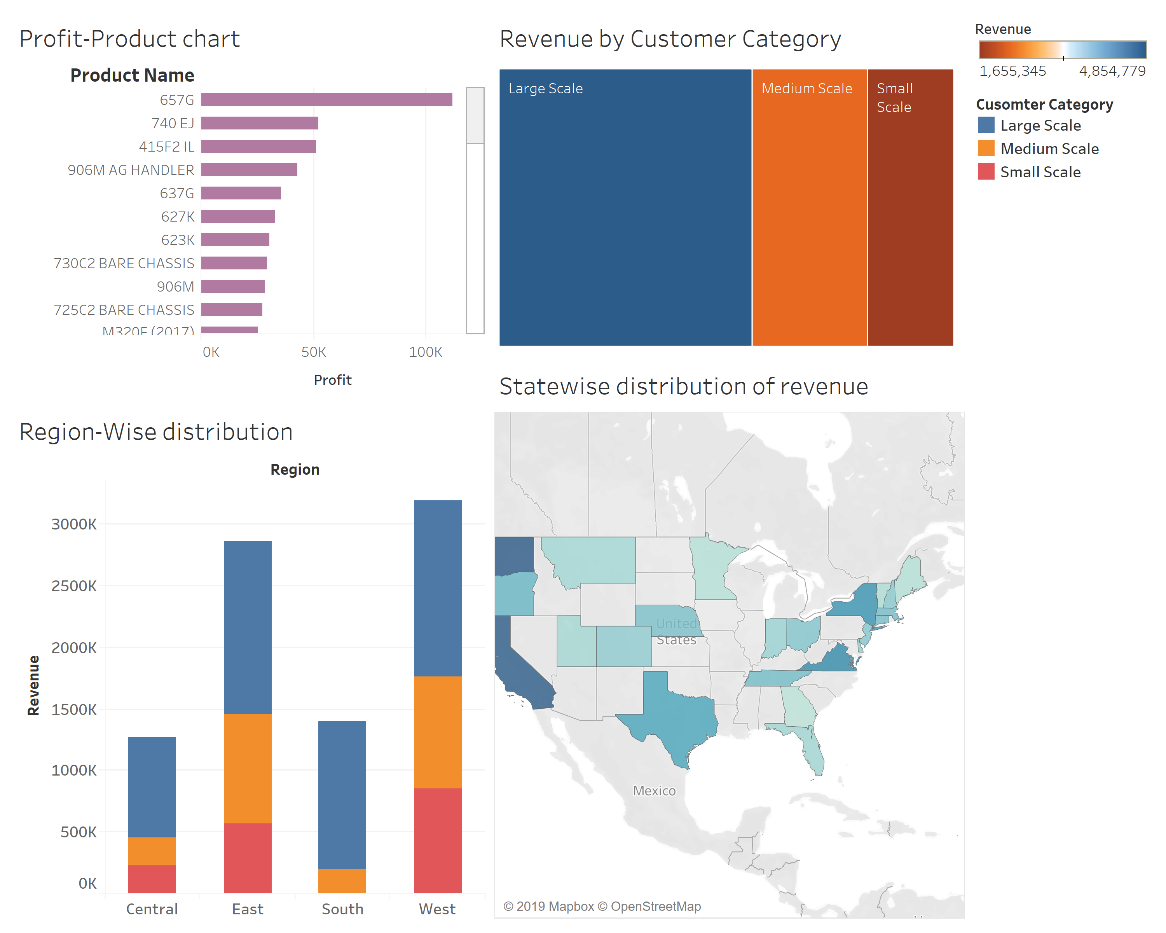
Below is the CSV file dataset which we have created.



**Approach:**

Once the data has been stored in the AWS Redshift database, we will be connecting the database to Tableau which is our preferred tool to create dashboards and provide solutions to our challenges faced in terms of analytics. We are also planning on using python to perform Exploratory Data Analysis (EDA) to get the basic idea of data as well.

Attached is a sample dashboard in tableau which we performed with the dataset we created:



**Connection to our database:**

We have created a group on AWS Redshift to give read access to users and in order for any other team’s member needs access to our database, we will be providing them with the following information and they would be able to read data from Redshift.

We would be needing the IP address a particular member of a team and we would add their IP address to the VPC security group and then they would be able to access the data by giving the following details below:

Host Name: info7374clusterassignment1team6.c76nrcvogw31.us-east-1.redshift.amazonaws.com

Port Number:5439

Database Name: info7374dbassignment1team6

Table Names: orders, customer, products

Username: team6

Password:Team6@dma

