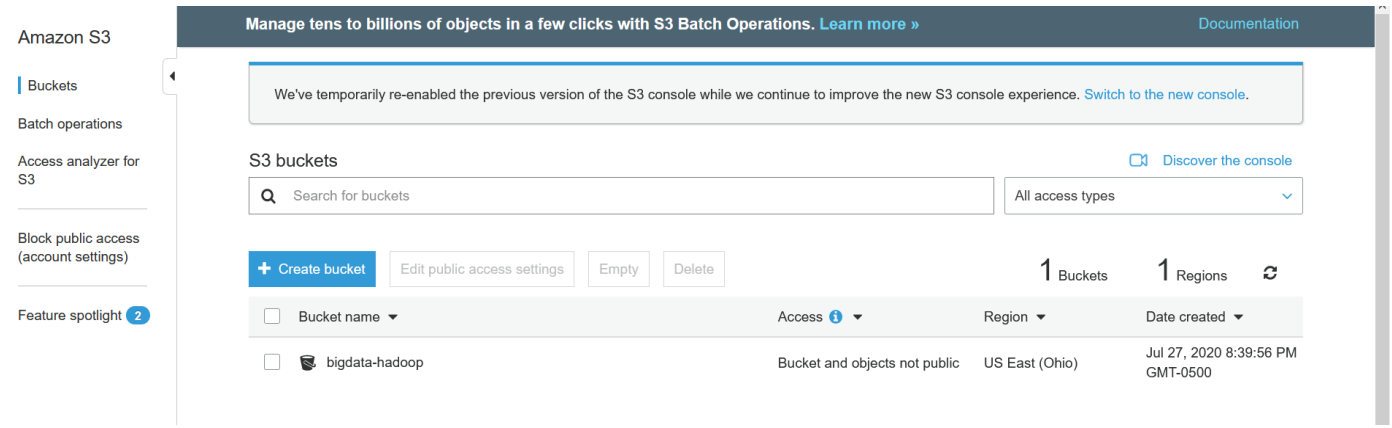
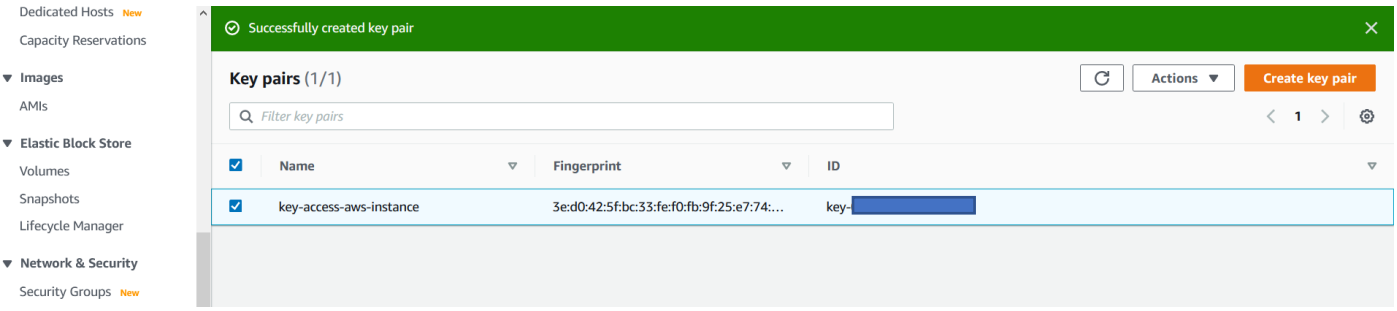
AWS Data Science

Analyze Big Data with Hadoop

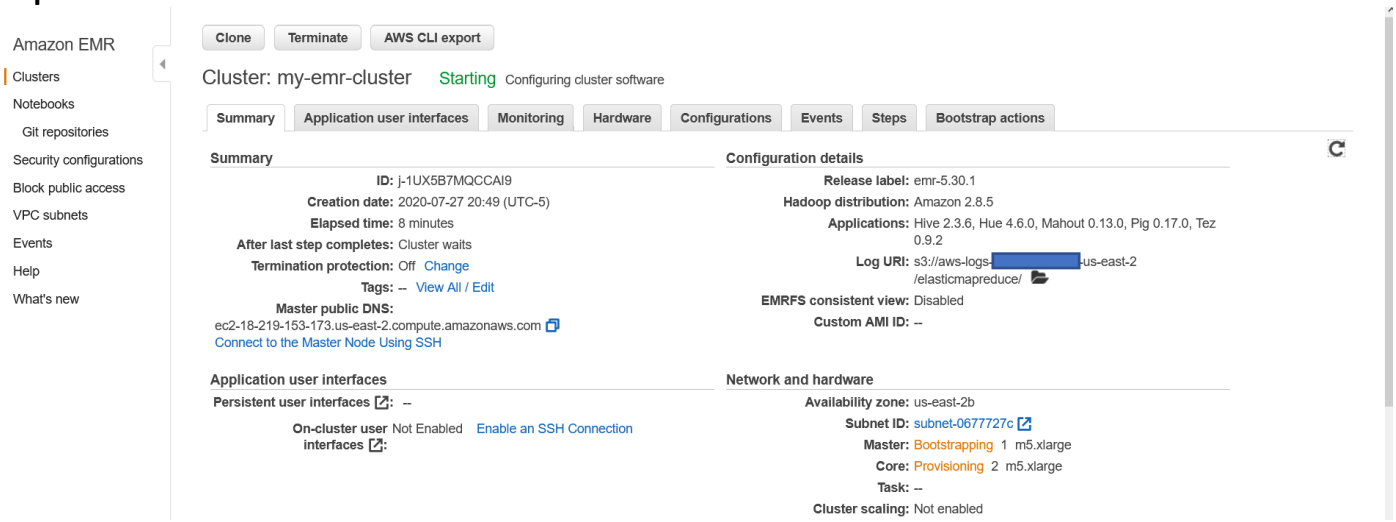
Create a Hadoop Cluster and run a Hive Script to process log data

Step1: Create S3 bucket and EC2 key pair

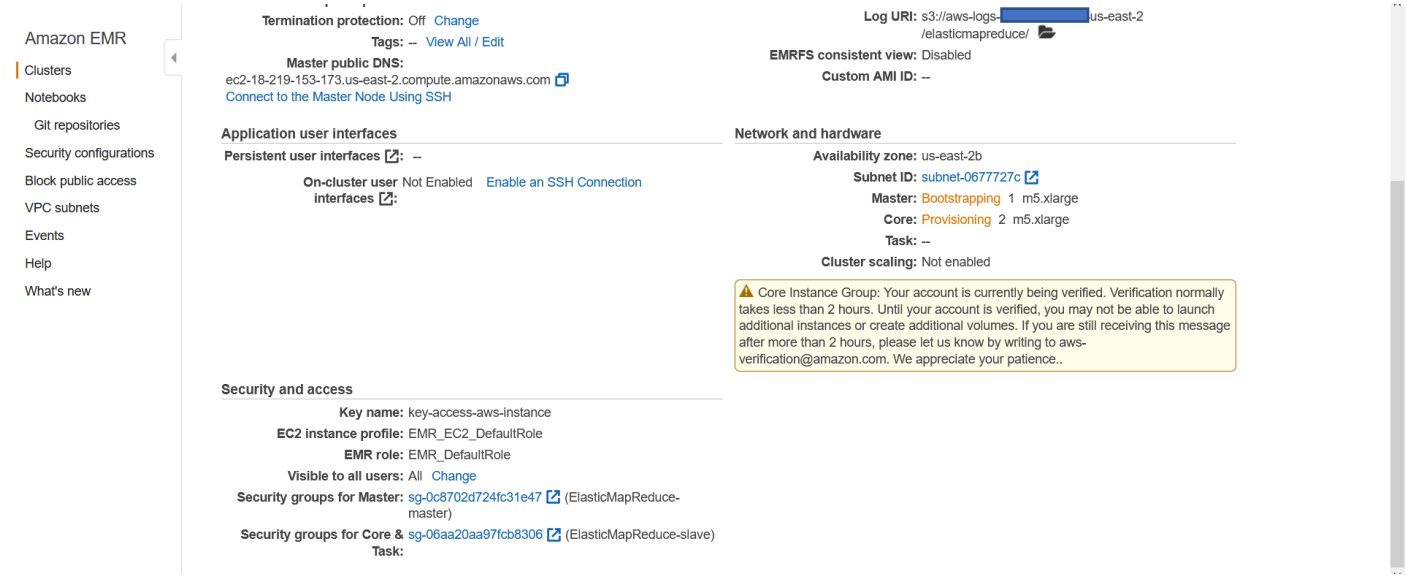


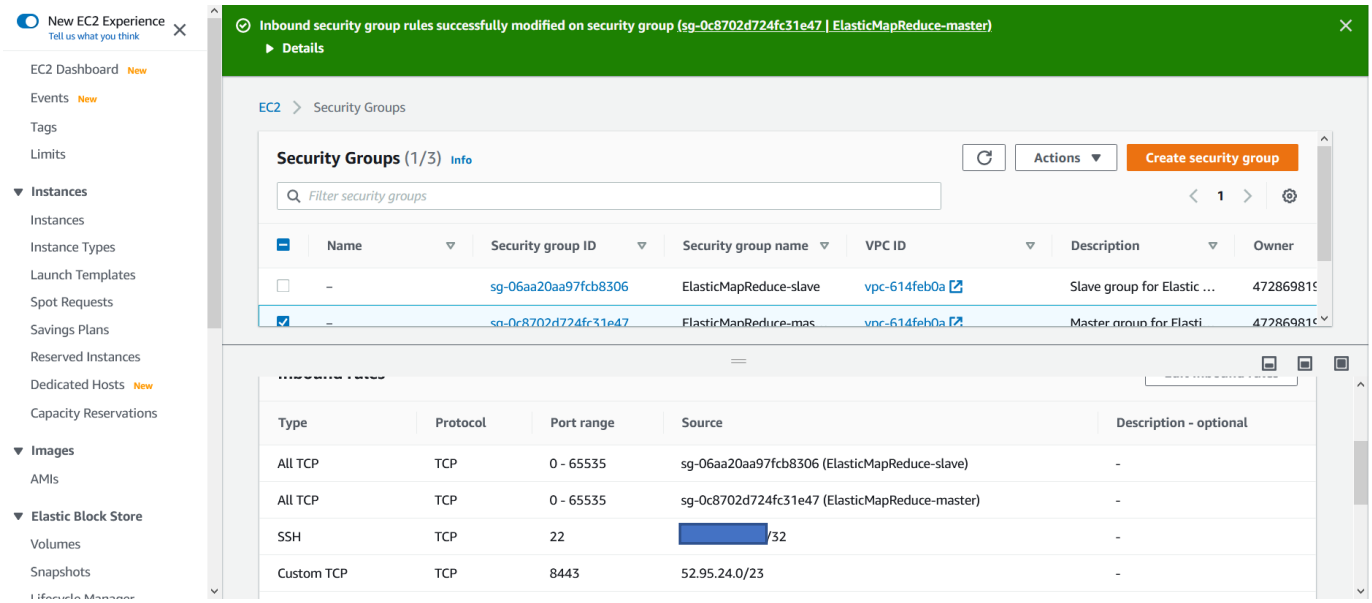


Step2: Launch Amazon EMR Cluster



Step3: Allow SSH access



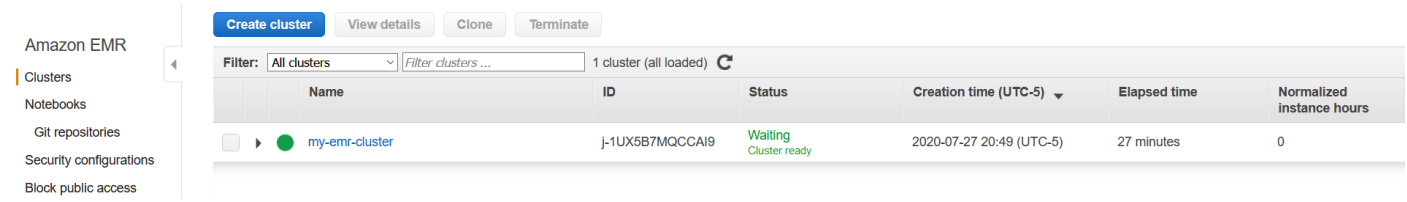


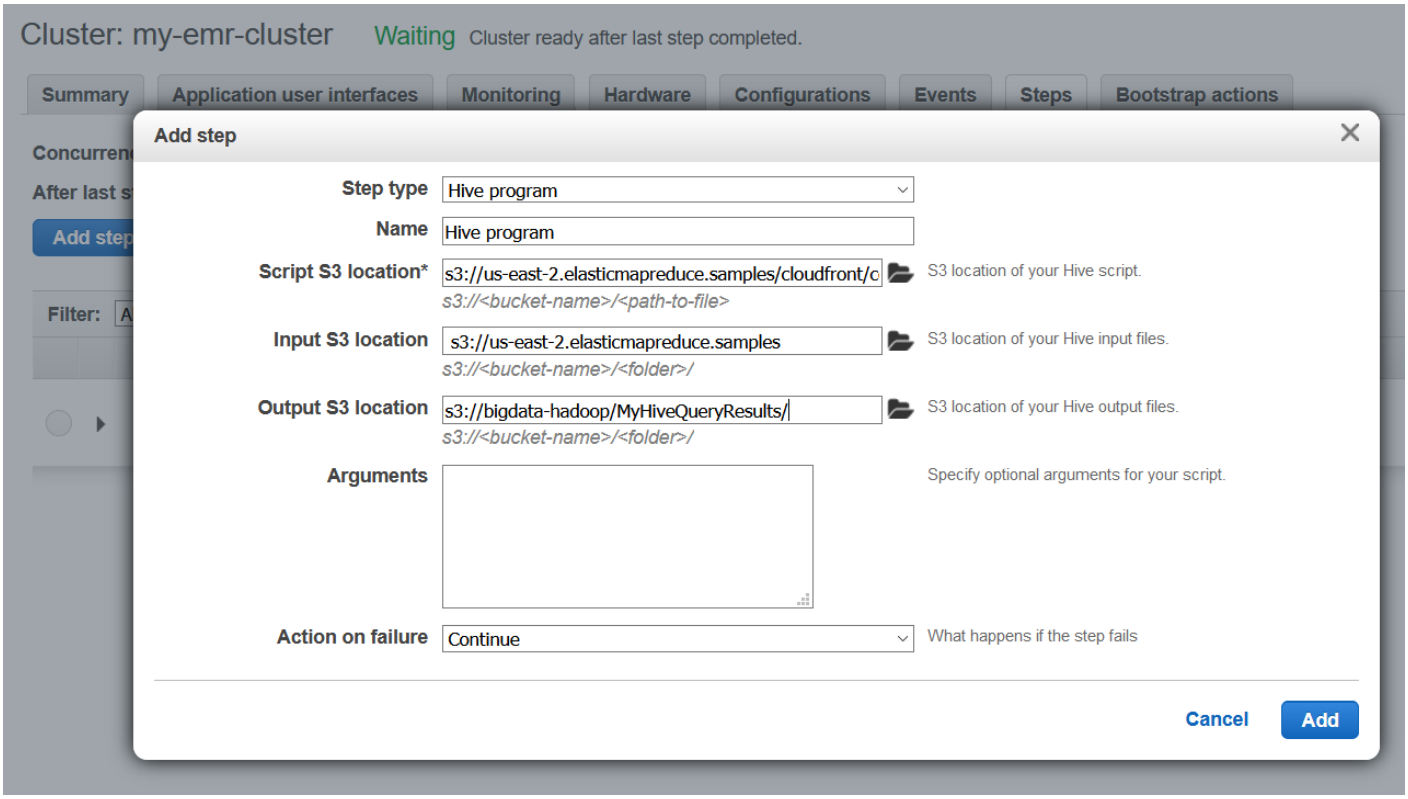
Step4: Run hive script to Process Data

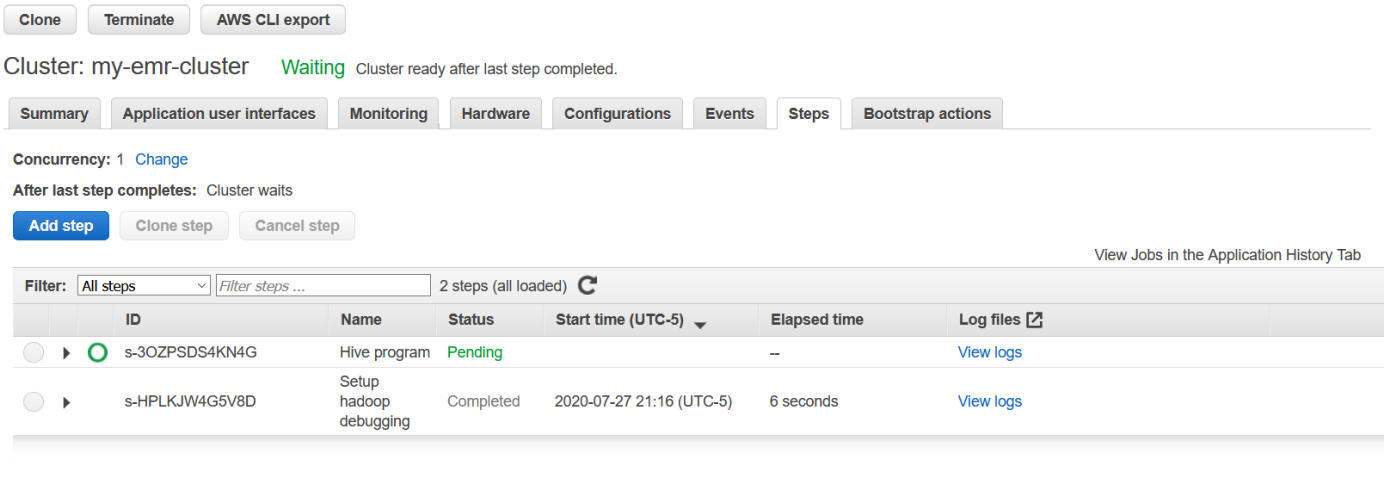
Hive_CloudFront.q

```
-- Summary: This sample shows you how to analyze CloudFront logs stored in S3 using Hive

-- Create table using sample data in S3.  Note: you can replace this S3 path with your own.
CREATE EXTERNAL TABLE IF NOT EXISTS cloudfront_logs (
  DateObject Date,
  LocalTime STRING,
  Location STRING,
  Bytes INT,
  RequestIP STRING,
  Method STRING,
  Host STRING,
  Uri STRING,
  Status INT,
  Referrer STRING,
  OS String,
  Browser String,
  BrowserVersion String
)
ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.RegexSerDe'
WITH SERDEPROPERTIES (
  "input.regex" = "^(?!#)([^ ]+)\\s+([^ ]+)\\s+([^ ]+)\\s+([^ ]+)\\s+([^ ]+)\\s+([^ ]+)\\s+([^ ]+)\\s+([^
]+)\\s+([^ ]+)\\s+[^\(]+[\(]([^\;]+).*\%20([^\/]+)[\/](.*)$"
) LOCATION '${INPUT}/cloudfront/data';

-- Total requests per operating system for a given time frame
INSERT OVERWRITE DIRECTORY '${OUTPUT}/os_requests/' SELECT os, COUNT(*) count FROM
cloudfront_logs WHERE dateobject BETWEEN '2014-07-05' AND '2014-08-05' GROUP BY os;
```

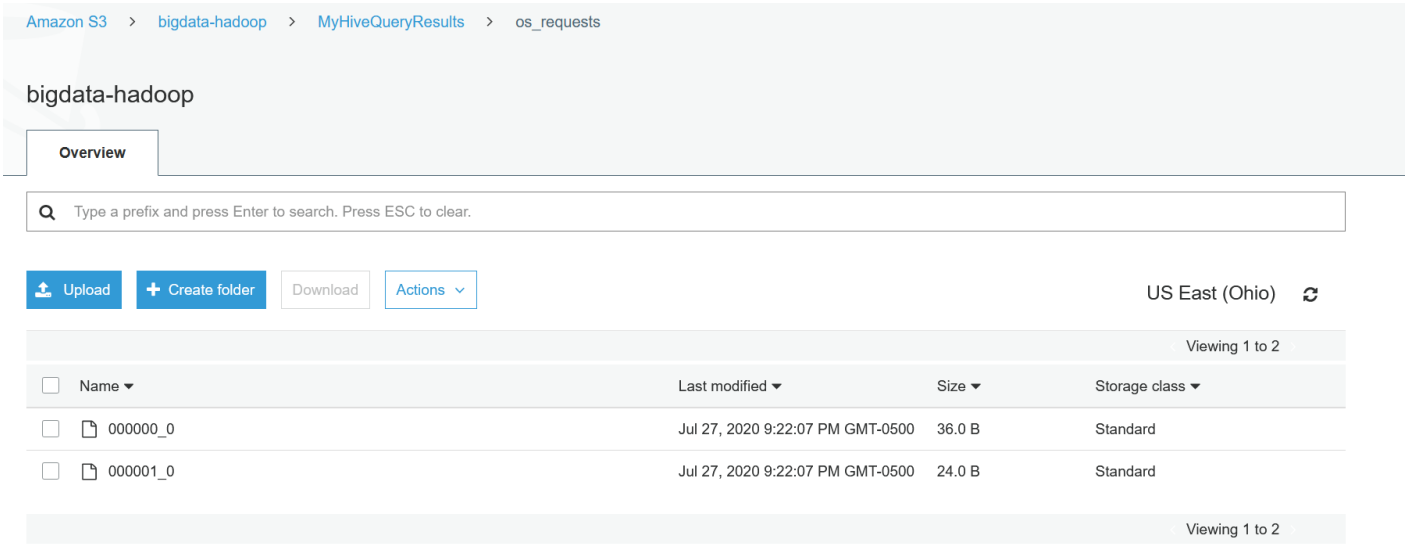
Step5: Submit the Hive Script as a Step

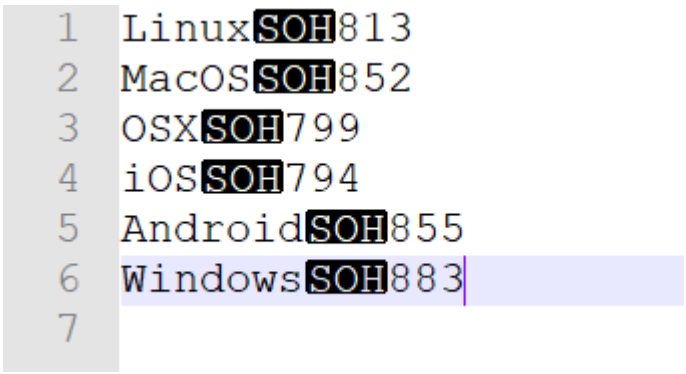




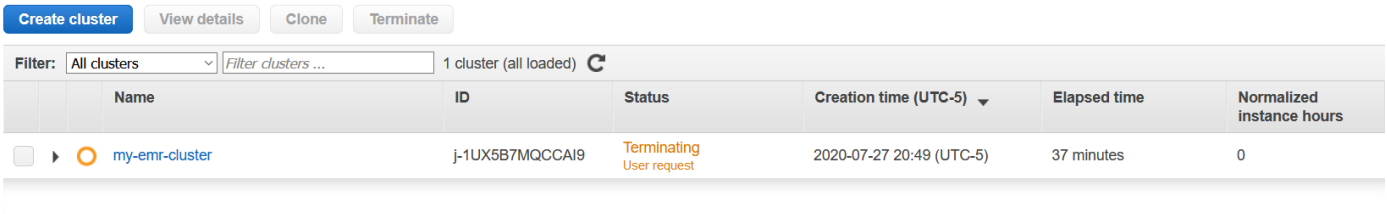


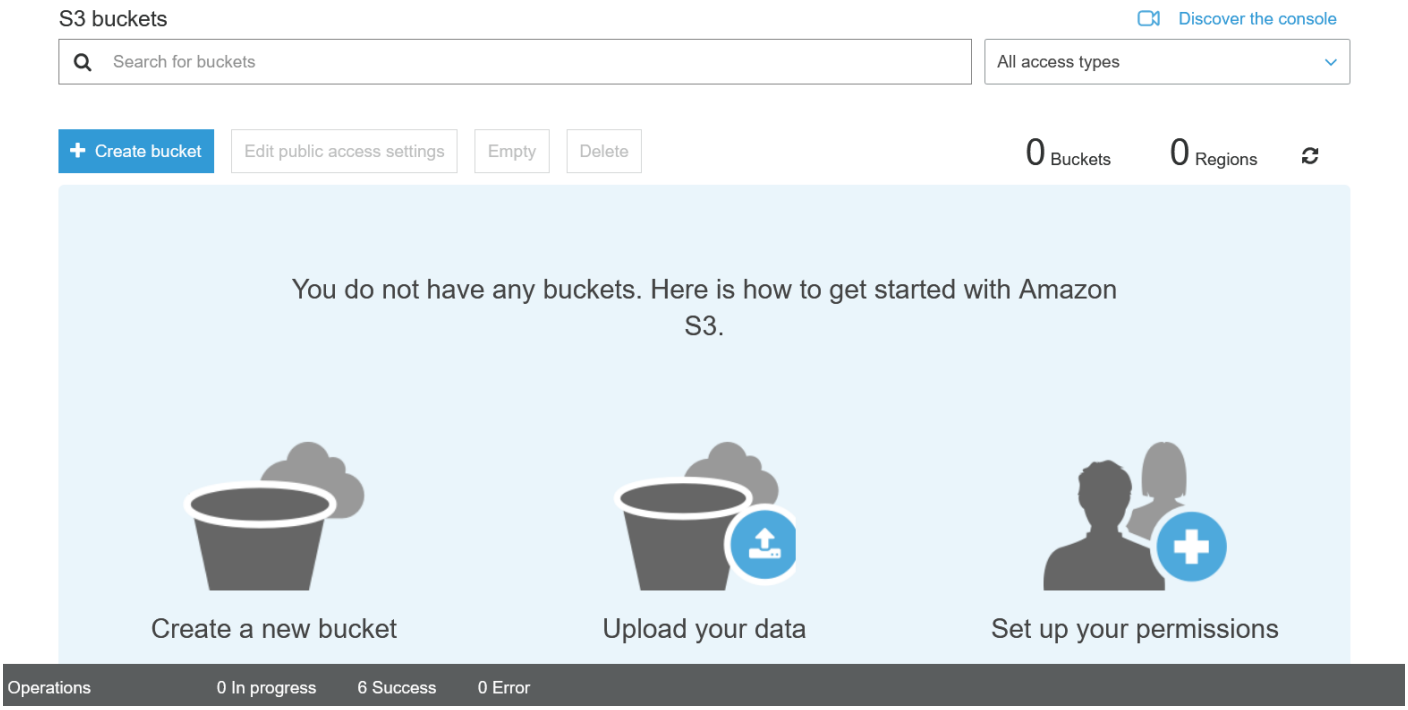
Step6: view the output of Hive script





Step7: Cleanup Resources




S3 buckets
Discover the console
Search for buckets
All access types
Create bucket
Edit public access settings
Empty
Delete
0 Buckets
0 Regions
You do not have any buckets. Here is how to get started with Amazon S3.
Create a new bucket
Upload your data
Set up your permissions
Operations
0 In progress
6 Success
0 Error