

**Data & Analytics**

**Homework #5**

March 2022



Data 228 - Big Data Technologies and Applications

Department of Applied Data Science

San Jose State University

***Faiza Ayoun (015960139)***

***Harsimran Kaur (016003468)***

***Pooja Malage (015294760)***

***Saranya Pandiaraj (015304497)***

# Table of Contents

[**Table of Contents**](#_heading=h.tvvzsxvr5j1y) **2**

[**Step 1: Set up Prerequisites**](#_heading=h.tyjcwt) **4**

[Setting up EMR](#_heading=h.3dy6vkm) 4

[Creating an s3 Bucket](#_heading=h.1t3h5sf) 5

[Uploading health\_violations.py and food\_establishment\_data.csv to s3 bucket](#_heading=h.4d34og8) 5

[**Step 2: Launch The Cluster**](#_heading=h.17dp8vu) **7**

[My Account id](#_heading=h.3rdcrjn) 8

[**Step 3: Allow SSH access**](#_heading=h.26in1rg) **9**

[**Step 4: Run a hive script to process data**](#_heading=h.35nkun2) **10**

[**Step 5: Clean up your Amazon EMR resources**](#_heading=h.1ksv4uv) **13**

[Terminating the Cluster](#_heading=h.44sinio) 13

[Deleting S3 resources](#_heading=h.2jxsxqh) 13

[References:](#_heading=h.z337ya) **15**

*This assignment is designed to walk you through the process of creating a sample Amazon EMR cluster using Quick Create options in the AWS Management Console. After you create the cluster, you submit a Hive script as a step to process sample data stored in Amazon Simple Storage Service (Amazon S3)*

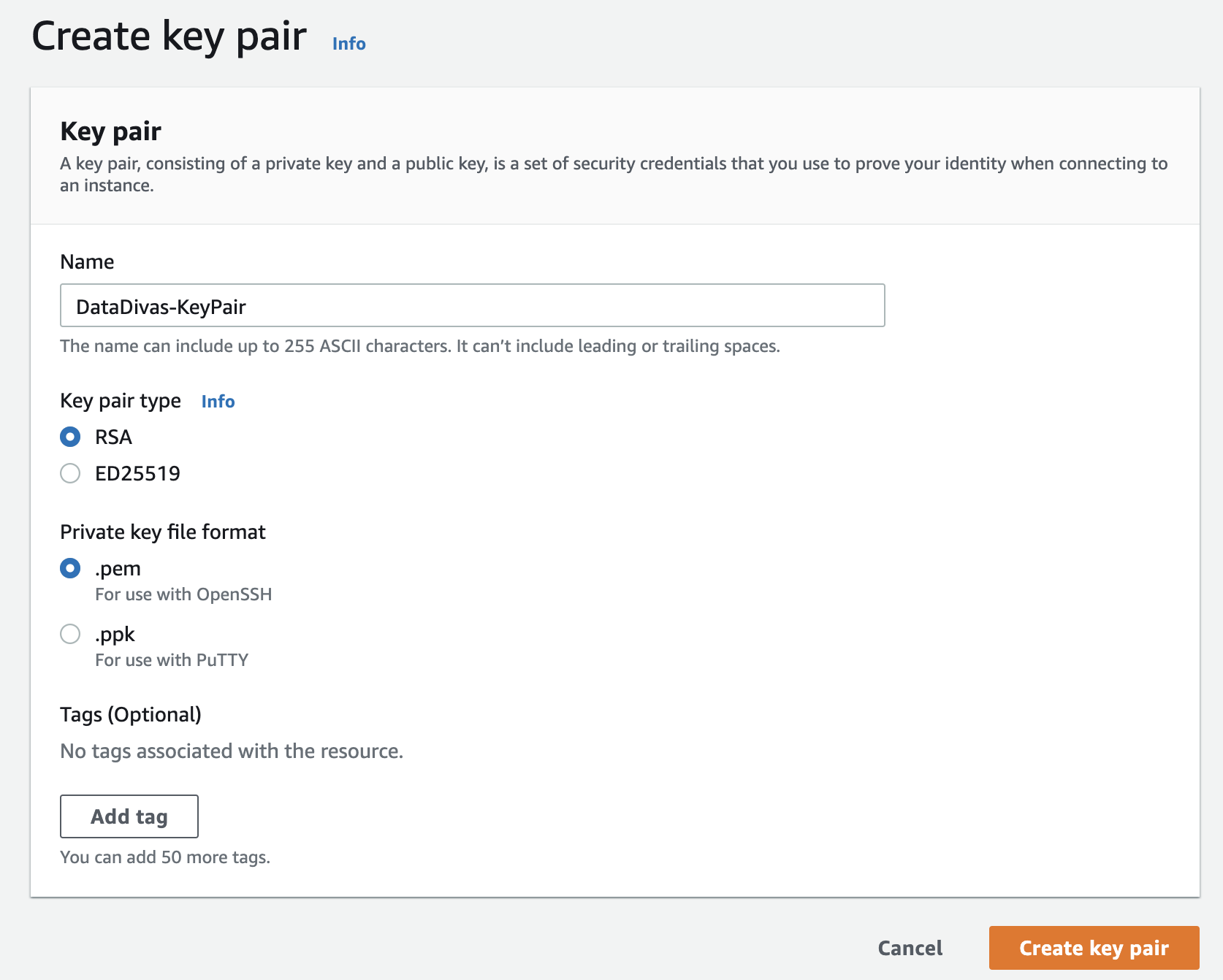
**Follow the steps and submit screenshots / report of your progress.**

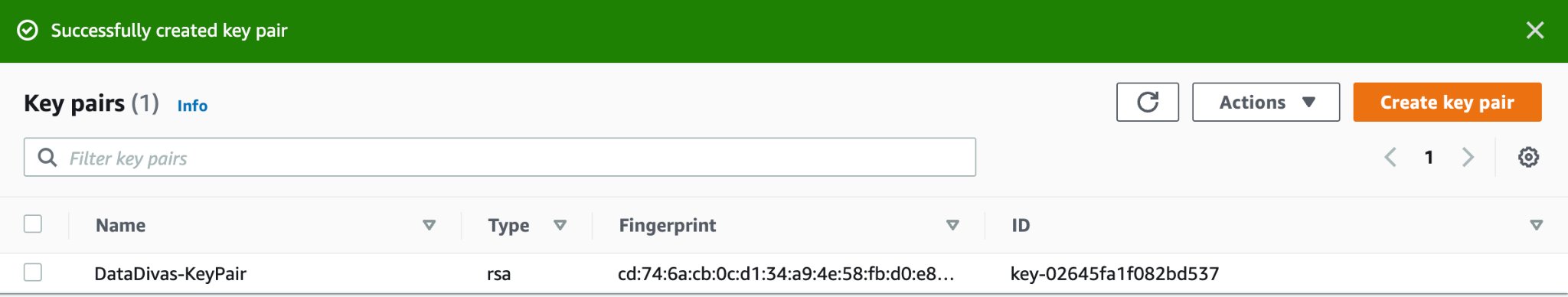
[**https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-gs.html**](https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-gs.html)

****

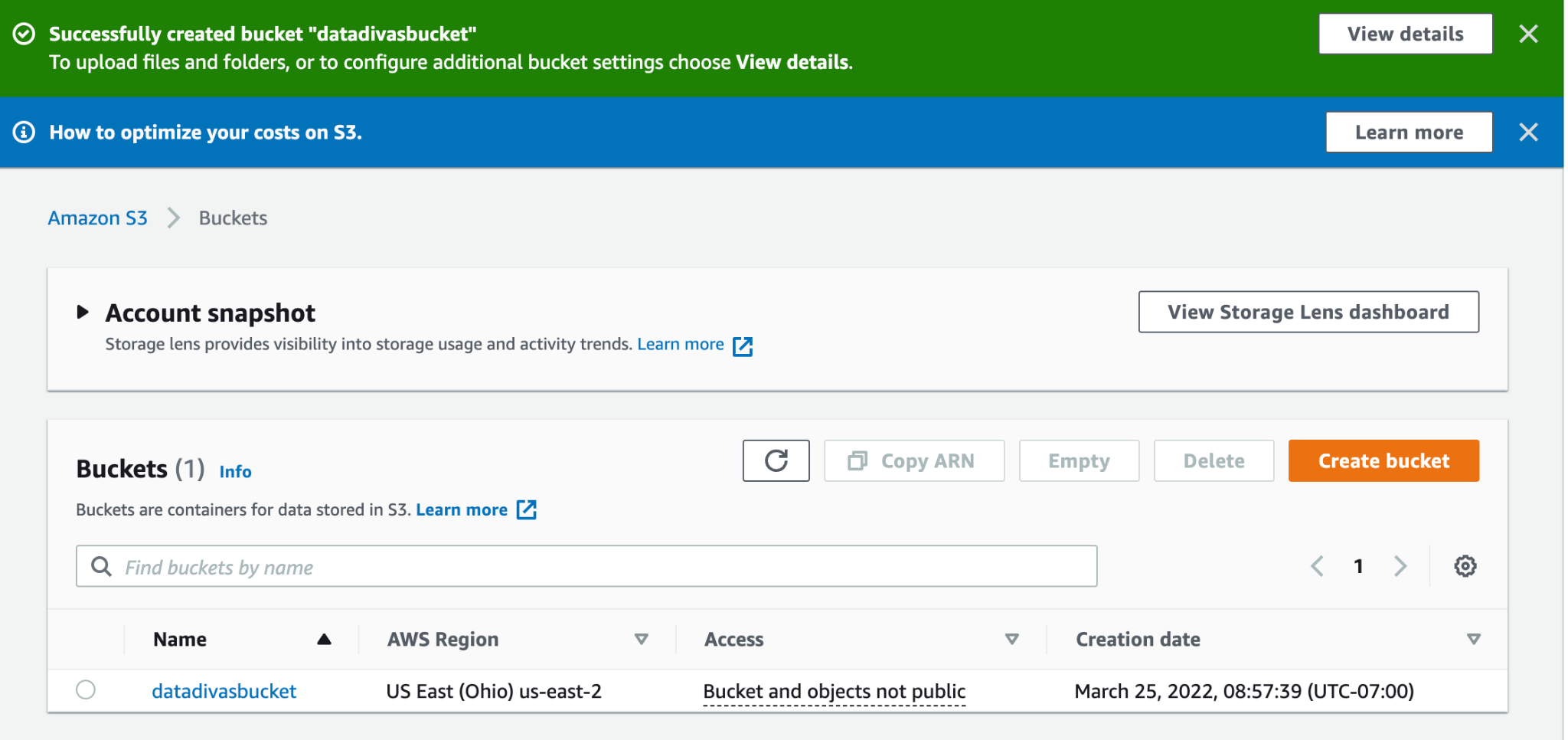
# Step 1: Set up Prerequisites

## Setting up EMR

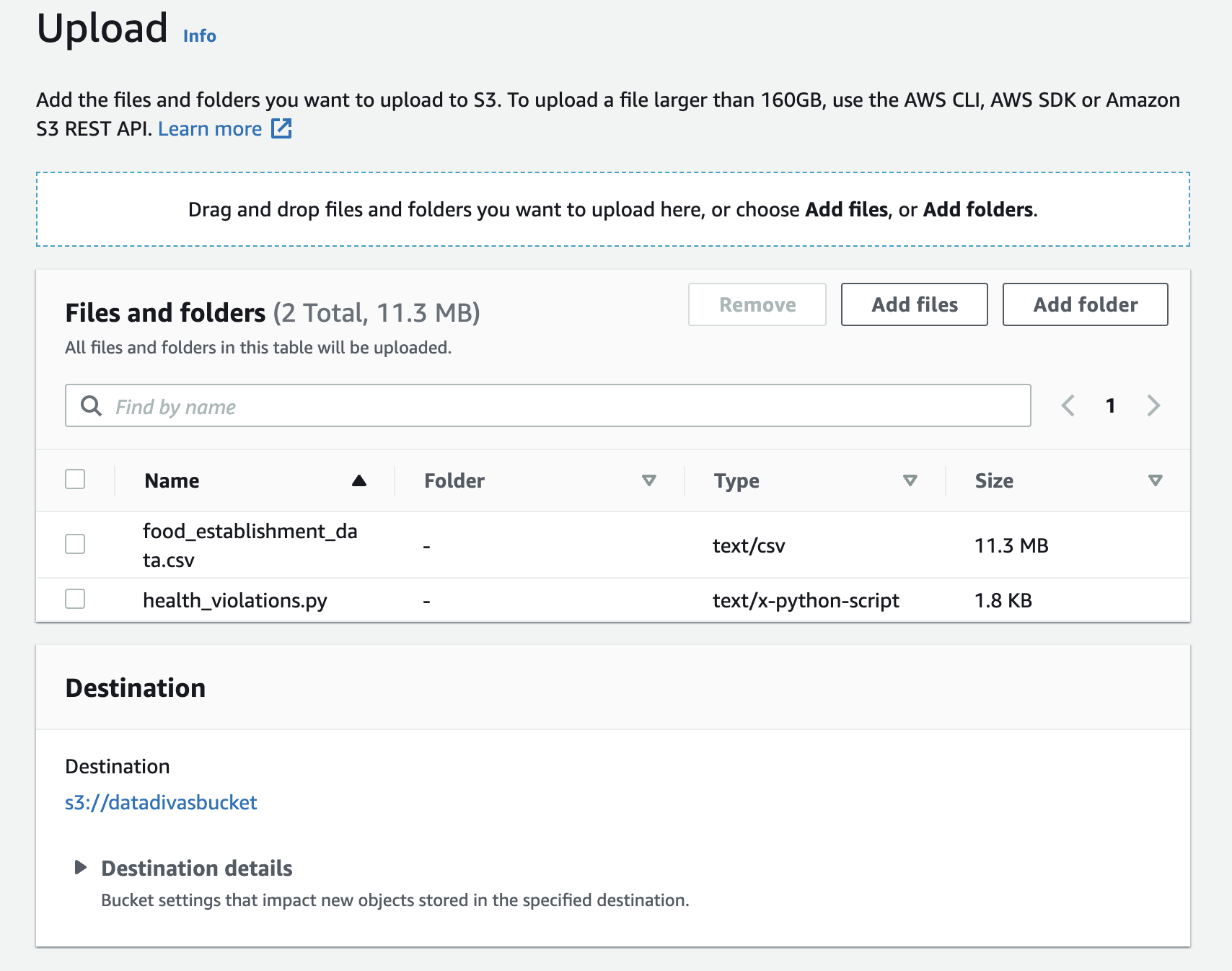
Create Key PA****

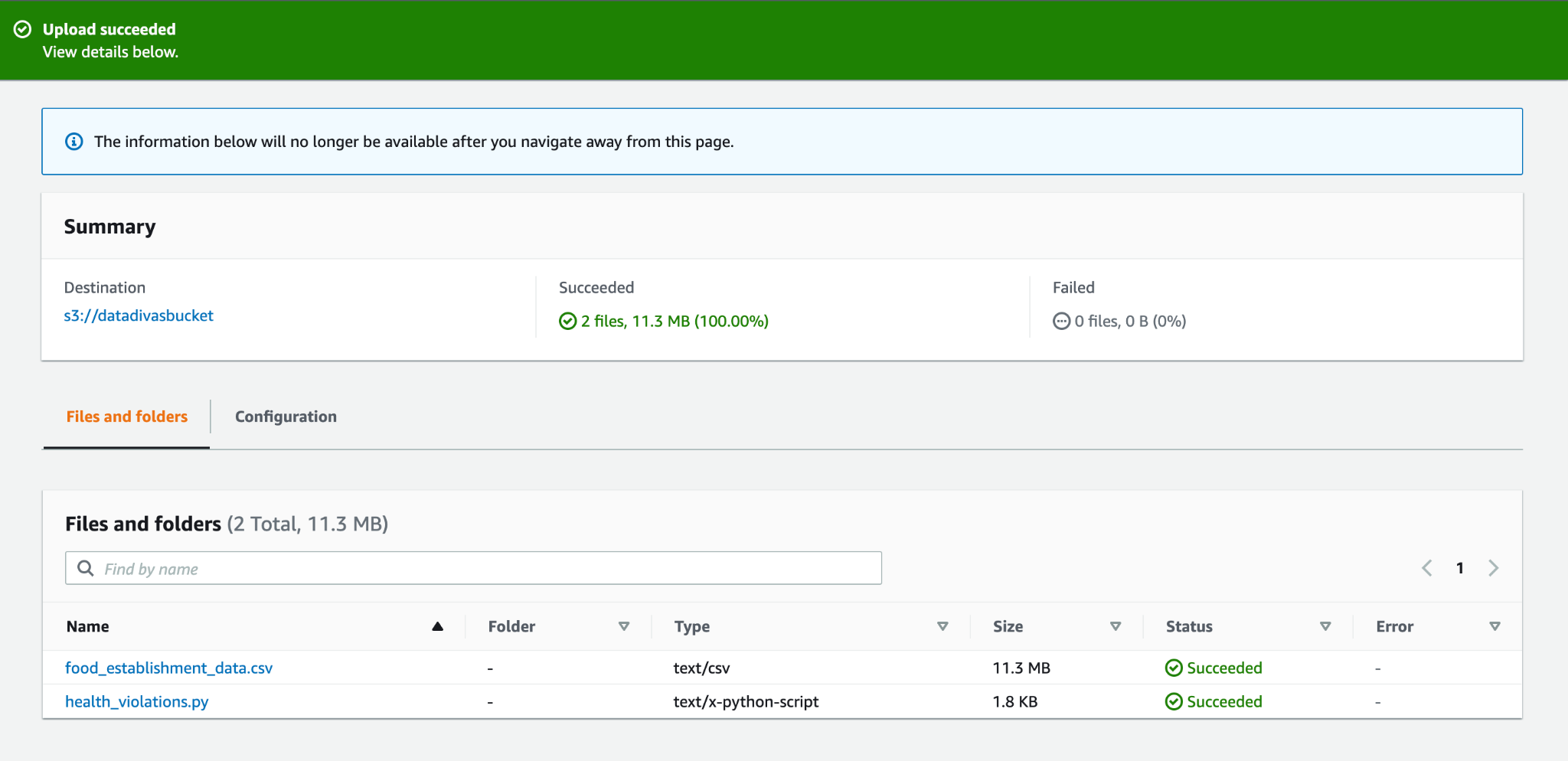
****

## Creating an s3 Bucket

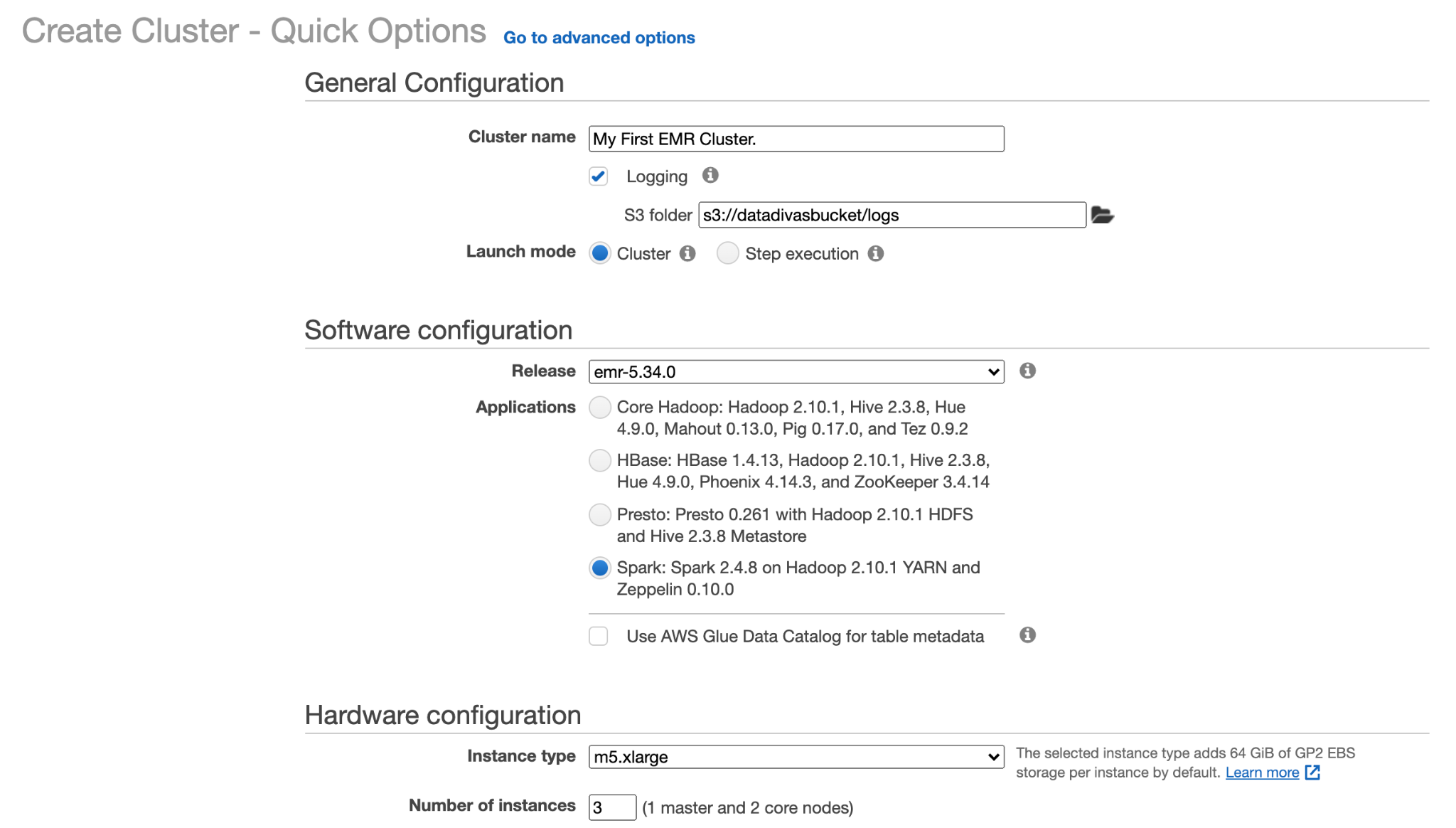


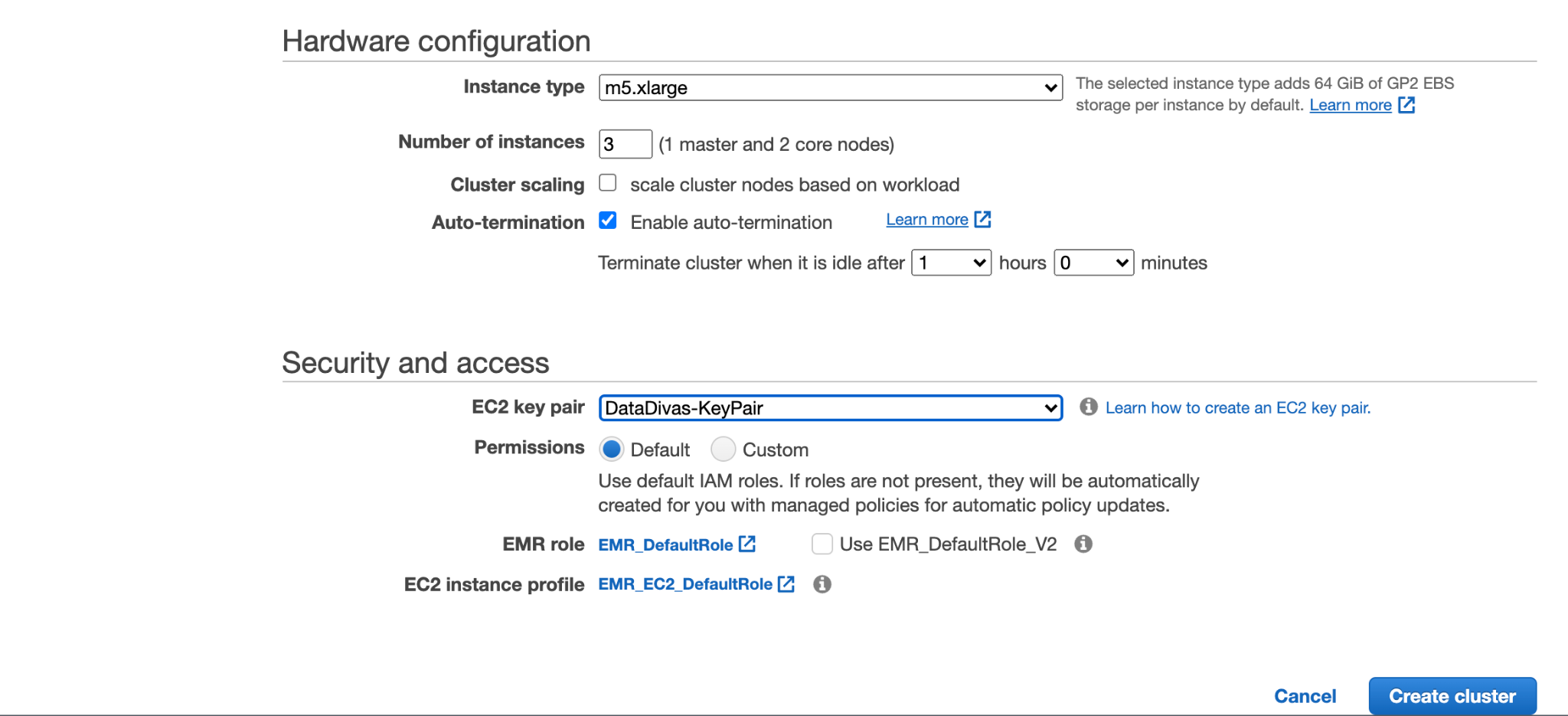
## Uploading health\_violations.py and food\_establishment\_data.csv to s3 bucket

****

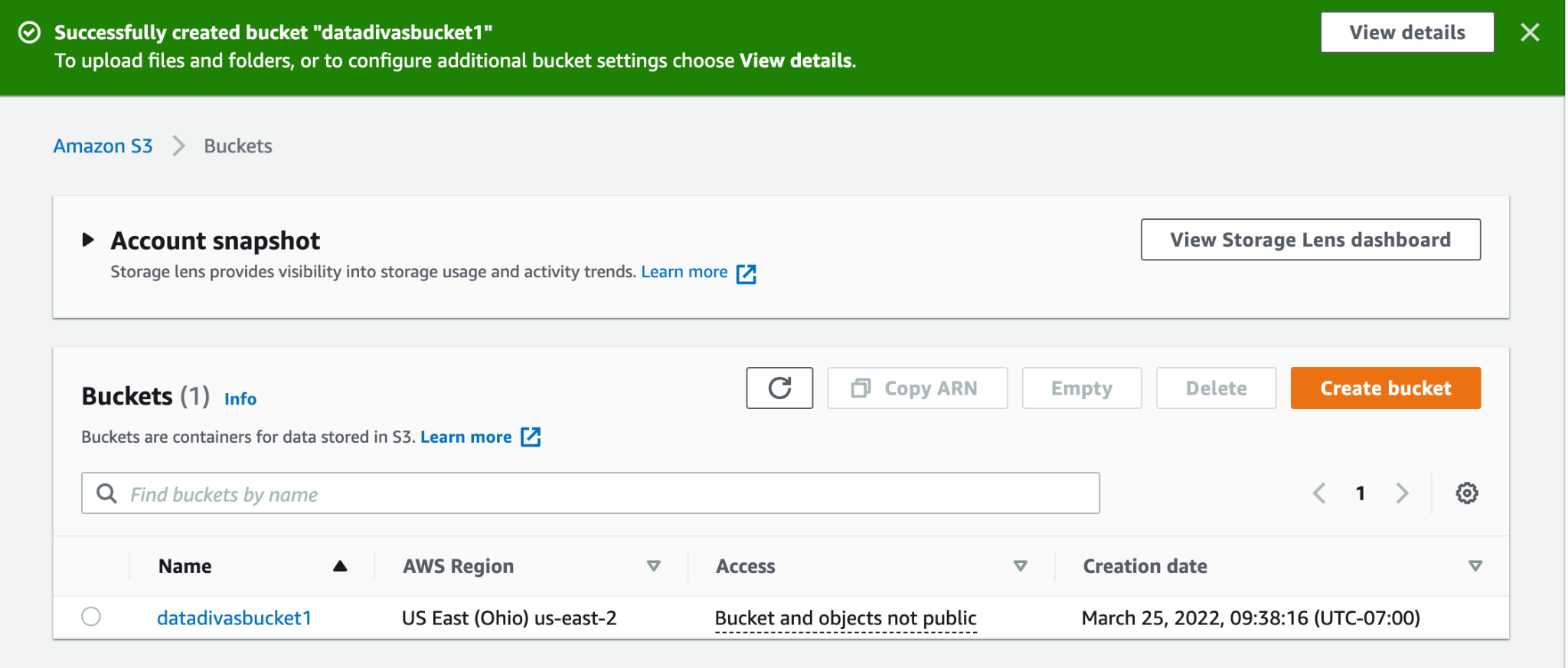
****

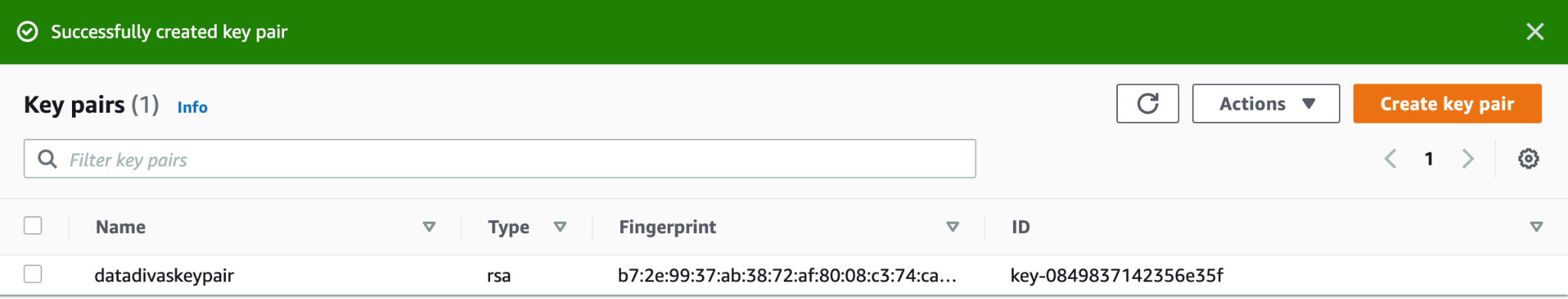
# Step 2: Launch The Cluster

****

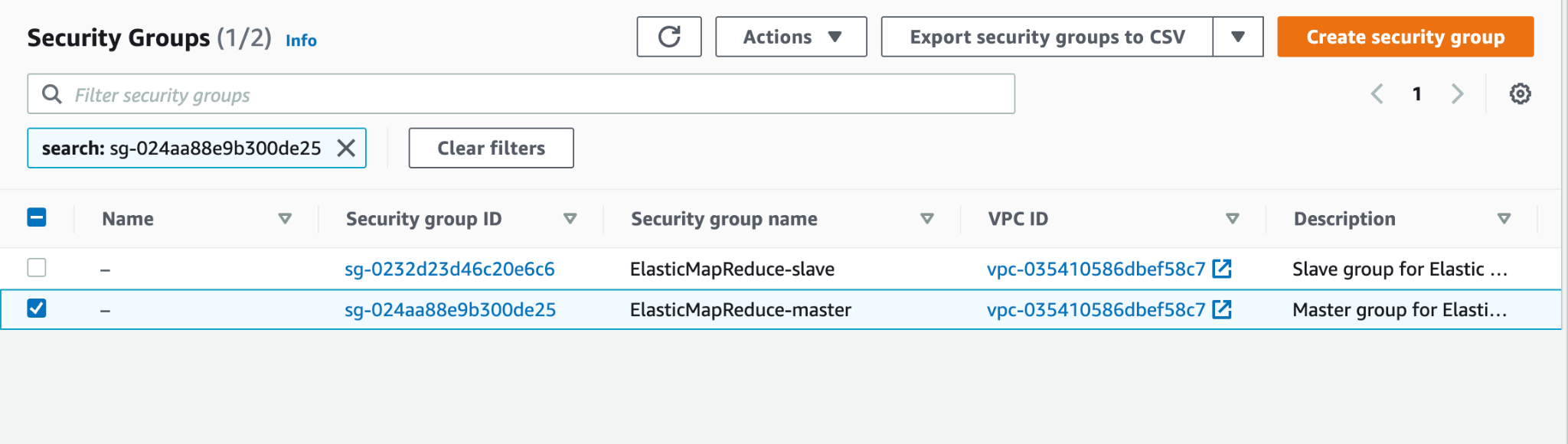
****

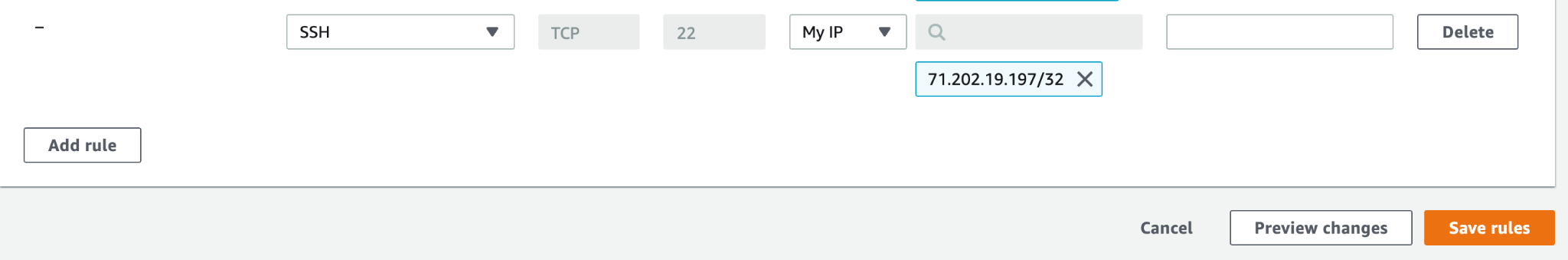
## My Account id



****

# Step 3: Allow SSH access

****

****

****

# Step 4: Ru[n a hive script to process data](https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-gs-process-sample-data.html)

Graphical user interface, text, application

Description automatically generated

**Graphical user interface, text, application, email

Description automatically generated**

**Graphical user interface, text, application, email

Description automatically generated**

**Graphical user interface, text, application, email

Description automatically generated**

**Graphical user interface, application, table

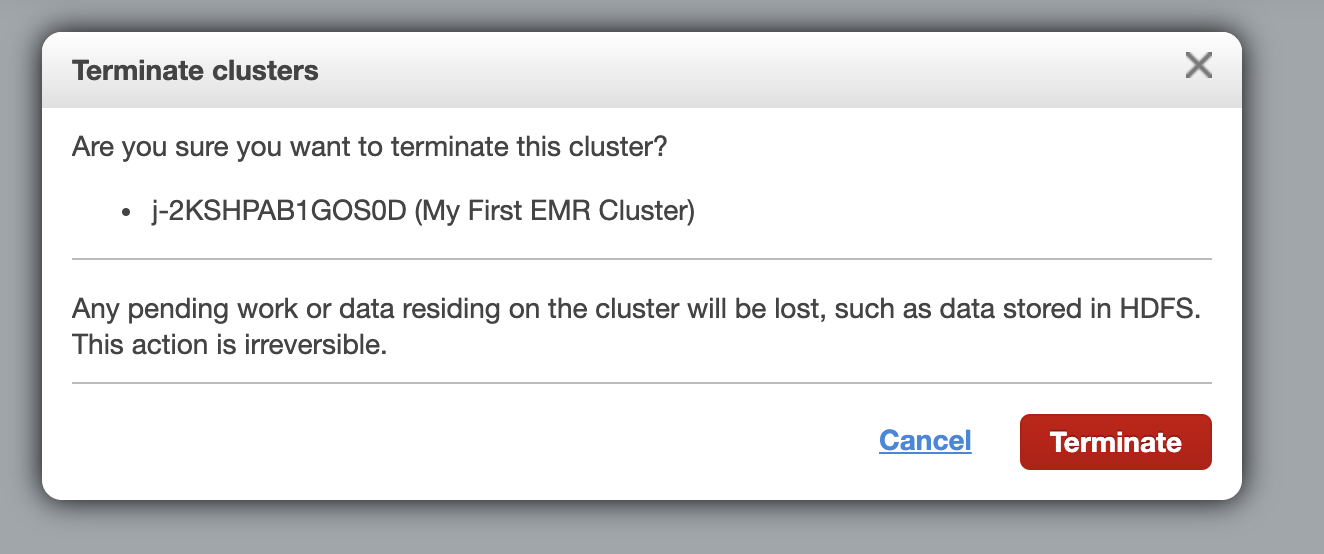
Description automatically generated**

**Graphical user interface, text, application, email

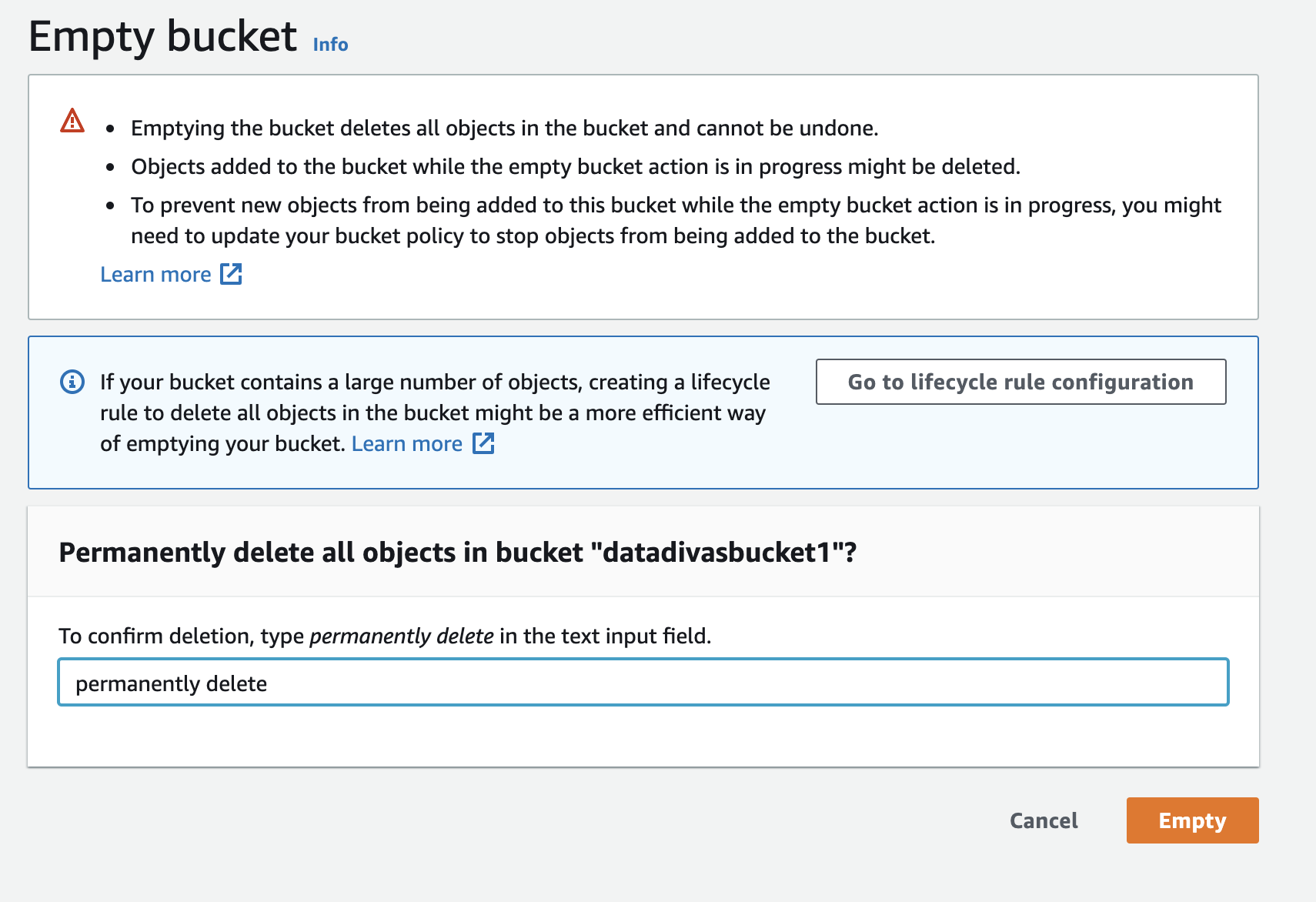
Description automatically generated**

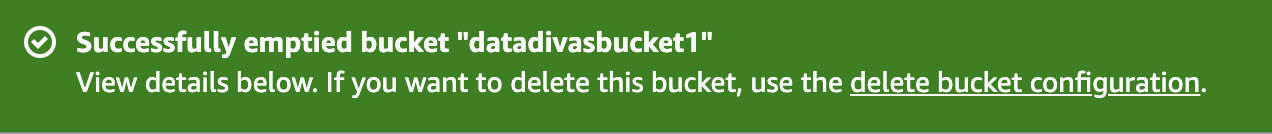
# Step 5: Clean up your Amazon EMR resources

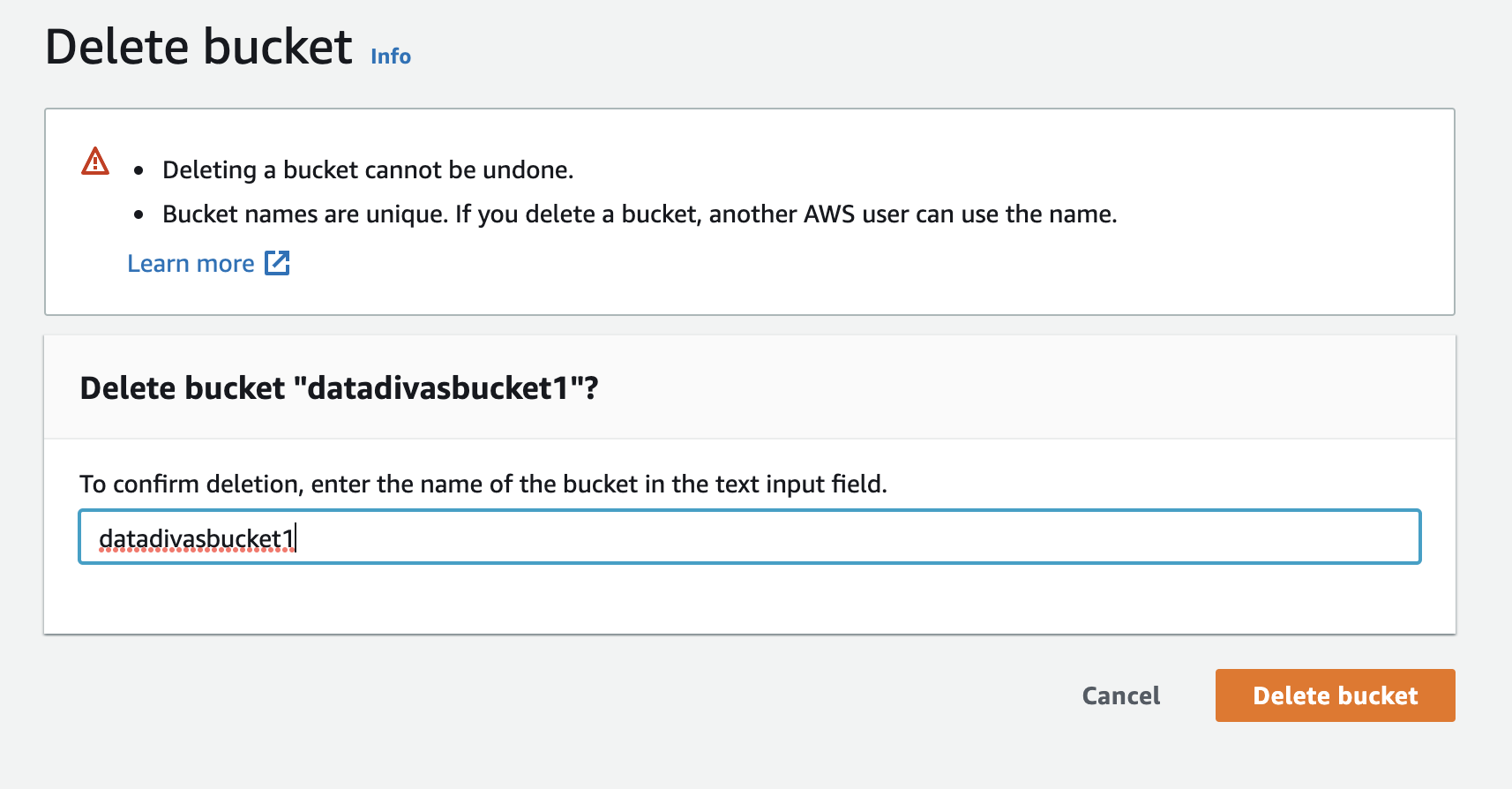
## Terminating the Cluster

****

## Deleting S3 resources

****

****

****

# References:

* (2022). Retrieved 2 April 2022, from https://docs.aws.amazon.com/emr/latest/ManagementGuide/emr-gs.html.