Prerequisites

I set up an S3 bucket called "seti-hive-test" with an input folder, output folder, and the scripts. I also set it as the target for logs, later, when setting up the cluster.

Graphical user interface, text, application, email

Description automatically generated

Launching the Cluster

I then launched a cluster called "CC-Practical". The logs target is set to the S3 above.

Graphical user interface, text, application

Description automatically generated

Setup SSH Access

The only requirement for this seems to be that an EC2 key-pair is assigned to the cluster on launch, which I did:

Table

Description automatically generated with medium confidence

Run a Hive Script to Process Data

This took 3 attempts. The first time I ran it gave me a syntax error:



It seems the script given had a mistake - the name of the field cannot match the name of a datatype. Renaming "Data" to "MyDate" fixed that problem.

Attempt 2 took 25 minutes to run! And then failed.

Graphical user interface, text, application, email

Description automatically generated

This time it was a mistake on my part, I forgot to change the output folder in the script. After fixing that, it ran successfully. This was on a second cluster the next day.

Graphical user interface, text, application

Description automatically generated

Cleanup Resources

After running the job successfully, I terminated the cluster.

Graphical user interface, text, application, chat or text message

Description automatically generated