# **Movie Search**

Search Engine for Movie Plot Summaries based on TFIDF and Cosine similarity

## **Upload files to Amazon S3**

* Create a bucket in Amazon S3

1. Upload the “moviesearch\_2.11-0.1.jar” to Amazon S3.

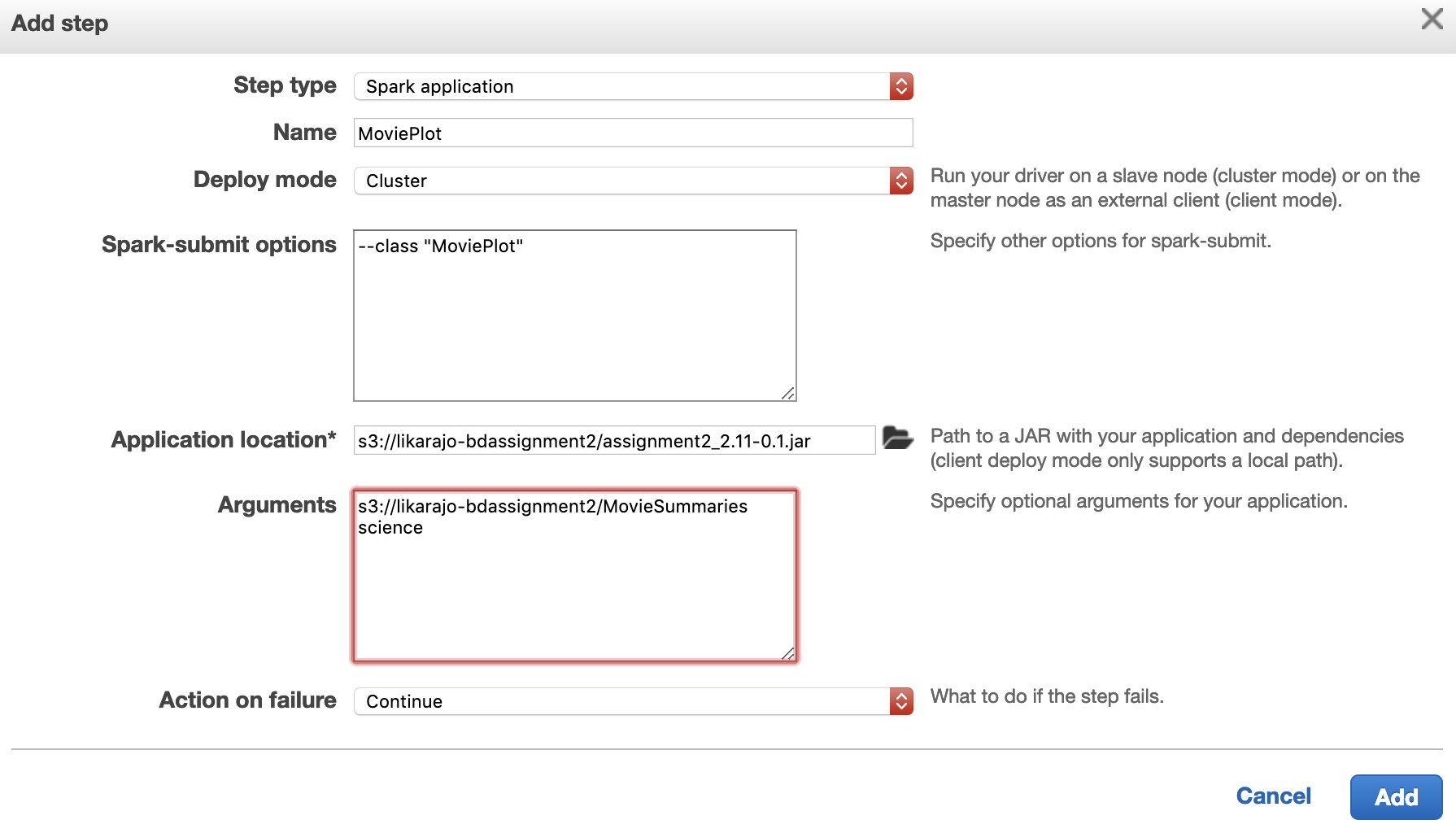
The jar file is inside the *target/scala-2.11* folder

1. Upload the “MovieSummaries” folder to S3 bucket at *s3://BUCKET/MovieSummaries*  
   The jar file is inside the *data* folder

## **Run the jar file in Amazon EMR**

* Create a cluster in Amazon EMR having Spark 2.4.0, and add steps to be run

1. Add Step to run class “MovieSearch” with arguments ‘input\_directory\_location’, and a ‘word’
2. Add Step to run class “MovieSearch2” with arguments ‘input\_directory\_location’, and multiple words separated by a space

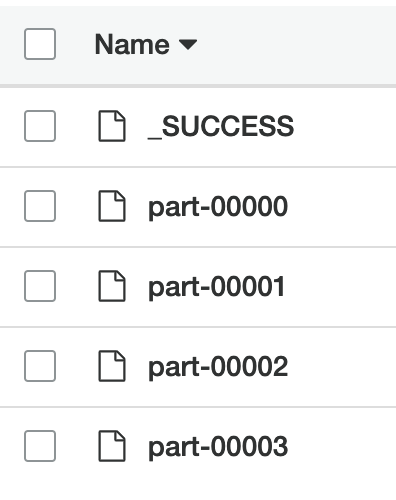
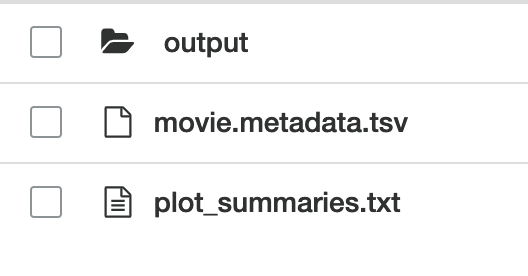


## **Verify the output files**

1. The output of MovieSearch or MovieSearch2 gets stored in specified output folder as well as ‘stdout’

A new folder named output is created under the specified data folder (passed as argument 1).

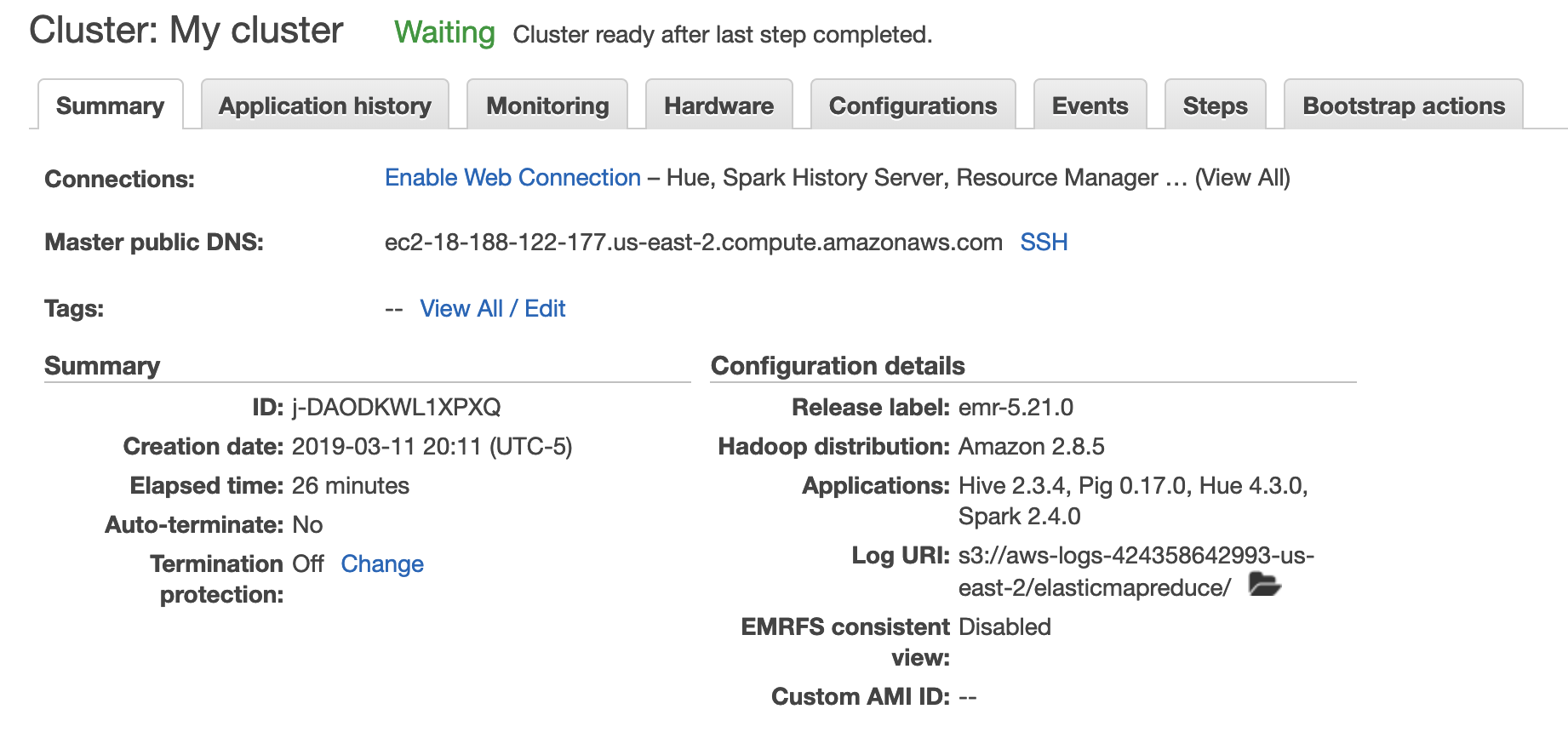
*s3://<bucket>/MovieSummaries/output*  
Inside the output folder are the generated part-0000x files.



The stdout files can be accessed here: *LogURI/containers/application\_id/container\_id/stdout.gz*

*Navigate to Log URI*

* *Navigate to Containers/*
* *Navigate to application\_id/*
* *Navigate to container\_id/*
* *View stdout.gz*



OUTPUT

