1. creating a sample Amazon EMR cluster
2. submit a Hive script as a step to process sample data stored in

In Amazon EMR, a step is a unit of work that contains one or more jobs.

The sample data is a series of Amazon CloudFront access log files. The data is stored in Amazon S3 at s3://region.elasticmapreduce.samples/cloudfront/data where region is your region, for example, us-west-2. When you enter the location when you submit the step, you omit the cloudfront/data portion because the script adds it.

Each entry in the CloudFront log files provides details about a single user request in the following format:

2014-07-05 20:00:00 LHR3 4260 10.0.0.15 GET eabcd12345678.cloudfront.net /test-image-1.jpeg 200 - Mozilla/5.0%2

The sample Hive script does the following:

* Creates a Hive table schema named cloudfront\_logs.
* Uses the built-in regular expression serializer/deserializer (RegEx SerDe) to parse the input data and apply the table schema.
* Runs a HiveQL query against the cloudfront\_logs table and writes the query results to the Amazon S3 output location that you specify.

Time spent: 30 min