收集EMR集群元数据

准备环境

1. 准备一台可以连接到EMR集群的EC2机器
2. 安装 Python3

安装数据收集程序并收集数据

1. 运行下面命令，安装EMR元数据收集程序

```
pip install emr-metadata-collector
```

2. 运行以下命令查看帮助：

```
emr-metadata-collector -h
```

注意：确保Python的bin目录在你的PATH环境变量里，否则需要使用安装包的绝对路径，比如"/home/hadoop/.local/bin/emr_metadata_collector"。

3. 运行以下命令列出正在运行的EMR集群，把{region}替换成集群所在的region, 比如us-west-2：

```
emr_metadata_collector list_clusters -r {region}
```

记录下所要收集数据EMR集群的ID。

4. 运行以下命令收集EMR集群的元数据，把{cluster_id}替换成上面记录下的集群ID, 把{region}替换成集群所在的region, {output_folder}替换成收集收据的目录，比如“emr_collected_data"：

```
emr_metadata_collector collect_data -c {cluster_id} -r {region} -o {output_folder}
```

5. 将上述命令生成的{output_folder}.zip 发送给AWS同事。