Automatic Job Scheduling of Big Data Jobs

1. **Purpose :-**

The purpose of this project is to automate the streaming jobs by a decider-worker model that integrates Amazon s3, Amazon EMR (Elastic MapReduce), Amazon SWF(Simple WorkFlow) by using python Boto3 library.
This model automates the daily/timely streaming/batch Jobs accordingly.
We need to register the workflow over AWS.

2. **Explanation :-**

The decider worker model is based on the following Algorithm steps :-

- Decider polls for a task. If task is available then it checks with worker if worker is sitting idle.
- If worker is idle, it asks worker to start a EMR Spark cluster for execution of the task with dynamically configurable settings. Then it takes a decision to assign the task to that worker by task_id.
- Worker responses with the acknowledgement id and starts waiting while the cluster gets ready to execute the jobs.
- After cluster is ready, worker submits the jobs to EMR Spark cluster.
- Worker always keep checks over cluster if everything is going fine. After Job Completion, the worker acknowledges the decider.
- Then decider again checks if any other jobs are available, if available it again repeat the above steps, else it terminates the cluster and go to a sleep of 30 minutes (it is configurable too).
- The sleep time for worker and decider are configurable.
- All the cluster settings are configurable.

Once this model starts execution, if goes on creating the supplied workflows. One workflow ends after 1 execution cycle is completed.