**Predator - An Experience Guided Configuration Optimizer for Hadoop MapReduce**

**Summary**

This paper introduces an experience guided optimizer, Predator, to optimize the execution time of Hadoop map-reduce job. When running Hadoop jobs, it is necessary to make full use of resources on cluster, which includes CPU, memory and I/O, optimizing Hadoop performance according to specific applications. The authors mentioned two ways to improve performance. One is changing the Hadoop core codes, but this is time consuming and should not be a general way. The other method is changing Hadoop parameters, which was considered as a better way and authors are working on. There are more than 200 configuration parameters on Hadoop and this paper studies 23 significant parameter of them. This paper categorizes these parameters into four groups. The basic idea of Predator is to preprocess the parameters within these four groups, analyzing the gathered information to estimate the time will cost to complete a job and then searching the optimal parameter value by experienced-combined search algorithm. The authors did experiments by running 1G, 5G, and 10G input data on 1 master node and 5 worker nodes. Jobs running with default values, configuration setting suggested by optimization based on random recursive search (RRS), optimization based on Predator and Configuration Model and Grid Hill Climbing algorithm (GHC-CM). Doing the experiments, the authors found that the Predator contributes to the highest efficiency and makes the local optimum close to the global optimum.

**Parameters List:**

1st group: merely experience

2nd group: consider cluster’s CPU and memory

3rd group: input info to be adjusted

4th group: search algorithm

A close up of a piece of paper

Description automatically generated