Programming on Cloud (Fall 2019-2020) - Assignment 2 Due

December 15th 23:55

Individual or Group of 2 Assignment

Problem Statement

This assignment aims to practice the MapReduce algorithm, its implementation and runtime.

The data set is from a Github project, under the directory of Workload Data.

https://github.com/haniehalipour/Online-Machine-Learning-for-Cloud-Resource-Provisioning-of-Microservice-Backend-Systems

The workload data contains the workload generated from two industrial benchmarks NDBench from Netflix and Dell DVD store from Dell. Both benchmarks are deployed on a cluster of cloud VMs on AWS and Azure clouds. The workload has been split to training sets and testing sets for machine learning purpose.

In each of the workload file, the first 4 columns contain the following attributes.

CPUUtilization_Average,NetworkIn_Average,NetworkOut_Average,MemoryUtilization_Average

In this assignment, we **only use the NDBench datasets**, training and testing.

Technical Requirements

1) develop a MapReduce program, given **one attribute name**, and **the data type** (training or testing) and 2) run the program in a MapReduce runtime on Hadoop, or AWS EMR, or MongoDB. Other MapReduce runtime should be confirmed with the lecturer; to 3) compute Maximum, Minimum, Median value and Standard Deviation of the selected attribute; and to 4) normalize all samples using feature scaling within the value of (0,1].

$$X' = a + \frac{(X - X_{\min})(b - a)}{X_{\max} - X_{\min}}.$$

You can choose to program the MapReduce implementation in Java, C++, Javascripts or Python. No matter what language you choose to implement, the submission should be packed and executable.

Individual Bonus Question (3 Marks added to the final)

Design your MapReduce program to produce the percentile 90% of the selected metrics. (Note: MapReduce may not be the most efficient method for a small size of data. This is only for practice purpose).

In your report clearly indicates the name of the author for this individual solution. If both members have solutions, it should be of two solutions.

Submission

Submit to Moodle site the following :
1. Pack all your source code and executable in a single zip file. .gz .tar or .zip are acceptable. Please do NOT use .rar file. The file should have this naming convention **[STUDENT1 ID_STUDENT2_ID]_A2_code.zip.**

2. A report in PDF with the naming convention **[STUDENT1_ID_STUDENT2_ID]_A2_report.pdf** that includes the following sections. The report should follow the format of IEEE publication. https://www.ieee.org/conferences_events/conferences/publishing/templates.html You can either use Word or Latex template. Make your report within 4 pages.

Section I. MapReduce Algorithm Design for technical requirements 3) and 4).
Section II. Instruction. How to run your program and show results with your sample data with screenshots.
Section III. Use the logs to estimate the time taken for the map task and reduce task respectively.
Section IV (optional) Bonus Solution.

Marking Criteria

1) Correctness of the program code, and outputs (30 Marks- Max 5 Marks, Min 5Marks, Median, 5 Marks, Standard Deviation 5 Marks, normalization 10 Marks)
2) MapReduce algorithm design (10 Marks – Section I in Report)
3) Quality of the report with clear description (10 Marks, Section II and III in Report)