Map Reduce Assignment 3 Report

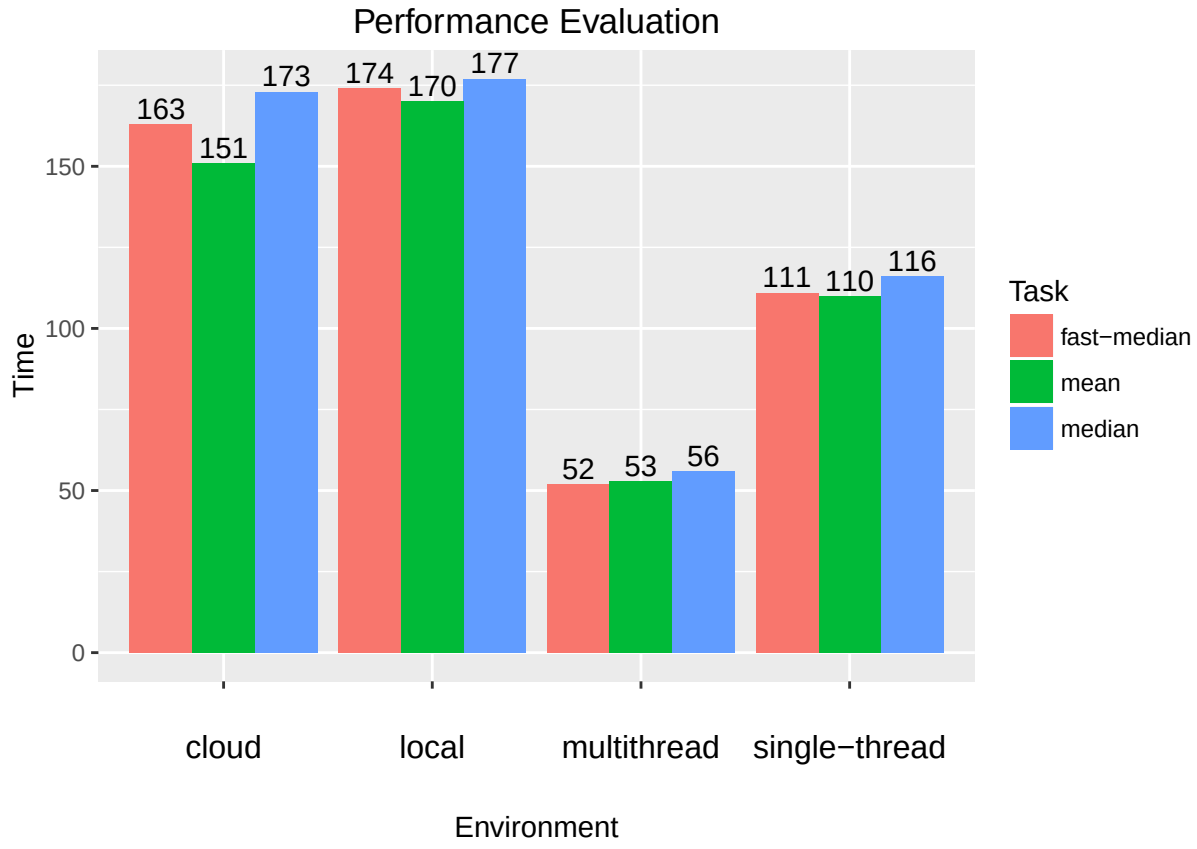
Authors: Akash Singh, Surekha Jadhwani

Description:

The aim of this assignment is to conduct benchmarking to compare the cost of computing (A) mean and (B) median price, and (C) fast median for different environments ((i) singe threaded Java, (ii) multi-threaded Java, (iii) pseudo-distributed MR, and (iv) distributed MR - AWS EMR) using the Bureau of Transport Statistics' On-time Performance (OTP) dataset which has over 27 years of air travel information about flights in the USA. In our program, sane records with Average ticket price missing or negative or higher than 100000 will not be used to calculate mean ticket price.

Our implementation of Fast Median uses approximation technique. It is 99.99% accurate with a deviation of 0.01%. For example: US carrier value for the month of December calculated using Median is 475.53 whereas using Fast Median is 470.35

The performance evaluation of different configurations using R script is as shown below:



Analysis:

We observed that single-threaded program was slower than multithreaded mode while multithreaded program (with 5 threds) had advantages over pseudo-distributed mode. We have not noticed advantages of using hadoop yet but may be with larger set of data, performance of hadoop will be better. With scaling of data, the fully-distributed cloud performance will improve as it was observed that cloud performance was comparatively better as the load increased (more input data files).