CodeReview for jashangeet-shreysa

Yang Xia

2017/10/29

Makefile: The code run after making the modifications the author mentioned in readme. And running on the sample data is pretty quick.

Report: When presenting the result, they include both id and actual name, which is good. There's no evidence supporting the conclusion. In the conclusion, the author says "while joining data points from 2 seperate files makes the program run very very slow", however, the comparision between runtime hasn't been shown. Having some data support the conclusion would be nice. And even nicer to have a pie plot showing the percentage time taken for each job. Also, author says "spark runs fairly fast", when they did not compare the result to hadoop implementation.

Code: line52: could have persist records to make the program run faster(when running on more than one nodes). line116: same problem, persist artisttuple. line18: why putting words like "version" in it?