CS 129. 1: Contemporary Database
Final Project Report
Carballo, Catolico, Mejia, Principe

Analyzing Lending Statistics Using MapReduce-based ANOVA

I. Big Data Problem

Given the dataset of loaners' information, the team aims to prove or disprove using Analysis of Variance if the means of the funded_income of the different types of loan_status are significantly different or not.

II. Data Source Description

The dataset has been retrieved from the lending and investing website called Lending Club. The dataset used for the project came from the year 2015 (only chosen arbitrarily). Impertinent columns that had 60-70% null values in it were removed using Microsoft Excel so that the size of the data is lessened. With this, it will be faster to apply MapReduce to the dataset because only the significant columns are included.

III. Results Description and Analysis

The output for this map-reduce function first yields the sum of all incomes per loan status classification (e.g. fully paid, current, etc.). Afterwards, on the one hand, the sum of all incomes per classification will be averaged, to get their own means. On the other hand, the sum of all the incomes of these loan statuses will be totaled to get the grand total, which will then be averaged to get the mean of the grand total. Each mean per loan status classification will then be subtracted from the grand mean. The values yielded from this computation will then be multiplied to the number of rows, which was what was yielded by the *count* function.

Each income will also be subtracted from their own individual means. Accordingly, the total of the sum of the squares will be retrieved computed. The quotient between the sum of the squares will then yield the ANOVA (Analysis of Variance) or the F-Statistic value.

Once this is gotten, the F-statistic table will be used to see if this F-stat value is within the 6 degrees of freedom. If it is, then the null hypothesis of the loan statuses having no differences will be rejected -- i.e., there are differences in the values of the income loan statuses from the given data set.

IV. Visualization of the Output