Statistical Analysis of Yield Data

You are going to write several MapReduce programs to perform statistical analysis of the semiconductor manufacturing process data. You can find the files from the following link:

http://archive.ics.uci.edu/ml/machine-learning-databases/secom/secom.data

This dataset contains 1,567 rows, each with 591 columns. 104 of 1,567 row represents yield failures. Here are the files:

- secom_labels.data - contains 1567 lines, each line has two columns: Yield (1 means fail, -1 means pass), and Date/Time of event
- Secom_date: contains 1567 lines, each line has 591 sensor readings.

Tasks:

- You can use any methods to write a MapReduce program to merge these two files together. As both files have the same number of lines. You can simply append the first line of one file to the first line of another file, and so on. After that, you will use the merged file for the following task:
- Write a MapReduce program to create a file that contains the records (lines) that failed the yield requirement.
- Write a MapReduce program to print yield fail (Yield = 1) counts by month:
    - 1   45
    - 2   0
    - 3   17
    - :
    - 12   88
- Write a MapReduce program to create a file that remove all sensor readings (columns) which have the constant values all 1,567 records, or NaN in any 1,567 records