Sarvender Dahiya

Insights

1. The original source files provided had 10 columns and variable rows of unclean data per file.
2. The data was cleaned by removing unwanted columns first.
3. Required new column ‘doctect/duration’ was created and every other column except ‘Real First Packet’, ‘doctet/Duration’ was dropped.
4. This process resulted in the reduction of combined size of source files from 566 MB to 56 MB of preprocessed files. Doing this speeded up the further calculations of the project.
5. Then more processing was done to arrange the average doctes/duration into time various time windows of 10s, 227s and 5 minutes from 8 am to 5 pm on weekdays for two weeks starting from Monday Feb 4, 2013.