

Data crunching

- We have obtained raw database from one of the popular site on the internet with large user base. Now we need to process that data and determine passwords of all users. The only issue is we don't have high end machines to process this data so we have to go with low end machines to perform this task.
- When we checked the raw data, it contained different TSV files (TSV files are same as CSV file with only difference of tabs instead of commas). The raw files are as follows
 1. User_email.hash.1m.tsv – 1 million records of users with id, username, email, hashed_password
 2. Ip_1m.tsv – 1 million records of users with id, ip
 3. Plain_32m.tsv – 32 million records of users with email, plaintext_password (Luckily!)
- Files are located at https://promactinfotech1-my.sharepoint.com/:f/g/personal/rushi_promactinfo_com/Eu2CDtrSu4tPhLPThGAKqsABb_pX448xyb8uJ8WnIPKgcA?e=U0cKV9
- Now we need to create tool/script/program which crunches data from these different TSV files and aggregates them to output file.
- The format of output file will be TSV with below fields
 - Id, username, email, hashed_password, plaintext_password, ip
- Limitations: We have low end machine so it can't run program which requires high memory and processing power. So, your code must not use memory more than 800MB.
- Submission: After completion, you can submit your approach and source code via email to me at rushi@promactinfo.com

Happy Coding!