**Hadoop Documentation**

4/8/2025

Number 3:

To begin the Hadoop processing job, the head command was run in Git Bash in order to see the columns of the .CSV file and pick which data I wanted to extract for Hadoop to process.

A computer screen shot of a computer error

AI-generated content may be incorrect.

NUMBER OF MOTORIST KILLED, and CONTRIBUTING FACTOR VEHICLE 1 where chosen to extract and use in the Hadoop file Processing.

Created a python program to clean the data even further and narrow down the columns to just NUMBER OF MOTORISTS INJURED, NUMBER OF MOTORISTS KILLED, CONTRIBUTING FACTOR VEHICLE 1, AND CONTRIBUTING FACTOR VEHICLE 2 for easier data processing in the Hadoop job.

After running Hadoop job, the results where different from what I expected. The leading cause of death in driving accidents is Unsafe Speed at 293 killed, and driver Inattention/Distraction in 2nd place 76. I was wrong about texting and driving being the leading cause of death by a long shot. And technically there isn’t a category for texting a driving. I am assuming it is in the Inattention/Distraction column.

My future plans with this data is to write a program to sort the data in order of most deaths to least so that I can see in order the leading cause of deaths while driving.