Scalable Data Mining: Assignment 1
Pankaj Mishra, 17EC35034

Question 1

Spark is used to load the data in an RDD instead of a list/array. The individual logs are stored as case class objects which makes matching records faster as the objects are immutable. In the first question, we only needed to count the number of records and filter. These operations can be made faster using Spark when using distributed systems.