Aufgabe 1 - ETL

1. Datenaufbereitung

Ein findiger Programmierer hat den Apache HTTPD server über ein Modul so angepasst dass neben normalen Seitenzugriffen auch mitgelogged wird um welchen Mitarbeiter es sich handelt, zu welcher Abteilung er gehört und welchen Kunden er gerade zugegriffen hat. Leider hat er sich keine grossen Gedanken gemacht wie ein Data Scientist die Daten verarbeitet.

Der Link zur Datei ist hier:
https://raw.githubusercontent.com/romeokienzler/developerWorks/master/log

Hier wurde einfach über ein Apache HTTPD modul für jeden Request eine 2. Zeile eingefügt in der der Payload die gewünschten Informationen enthält, in folgerner Reihenfolge: departmentid, employeeid, clientid

a) Lesen Sie die LOG Datei mittels R ein und bereiten Sie so auf, dass daraus ein Data Frame entsteht welcher folgendes Format hat: Spalte 1> employeeid, Spalte 2> departmentid, Spalte 3> clientid

b) Erweitern Sie Ihr R Script dass nun auch die Stunde des Zugriffsdatums aus der LOG Datei in der ersten Zeile des Data Frame erscheint. Das Format ist nun Spalte 1 > hour, Spalte 2> employeeid, Spalte 3> departmentid, Spalte 4> clientid

Der Link zur Datei ist hier:
https://raw.githubusercontent.com/romeokienzler/developerWorks/master/testdata.csv

2. Die bekommen nun das aus Aufgabe 1 extrahierte CSV file von Ihrem Junior Data Scientist geliefert. Die Forensik Abteilung möchte wissen ob in diesem Trace anomales Verhalten auftritt. Können Sie helfen?