PerimeterX - Home Assignment

Your mission, should you wish to accept it, is to create a data processing pipeline using Spark / classic MapReduce/ Hive.

You can find the raw input data in the following file:
https://drive.google.com/a/perimeterx.com/file/d/0B9ICLDMRLmwdRFhnNUFxWTZlenc/view?usp=sharing
The file contains dates and author names of public open source projects commits hosted on *github*.
Please, keep in mind, that your code should be able to handle significantly larger amounts of input data than contained in this file.

Tasks:

1. Calculate the average, standard deviation and 95th percentile of the number of commits per day-of-week. Print those results to the results log.
2. Find anomalous days, these are days where the number of commits is higher by at least two standard deviations from the average. Print those dates to your results log.
3. From the anomalous days you found in the previous task, take the day with the maximum number of commits and find the user who is responsible for this anomaly. Print his name to the results log.
4. *BONUS*: Can you explain what happened there? Who is this guy and what did he do?

How to submit your results:

1. You should send us the results log of the tasks above.
2. Please attach a short description of the data pipeline you have used to accomplish the tasks above. The description should contain each processing step you did on the data (e.g. loading, filtering, mapping, caching, aggregating, reducing, etc.).
3. You should also submit your code and all related files, together with a clear README on how to reproduce all the required environment to run the code. It should be reproducible on Ubuntu or Mac OS environment.