

DBS

# CA04 – Programming for Big Data

Processing a log file using python

# Contents

Introduction.....	3
The Basics .....	3
Figure 1: Commits per <b>author</b> .....	3
Figure 2: Comments per author .....	3
Figure 3: Commits per week.....	4
Figure 4: Commits per day.....	5

## Introduction

A GitHub logfile was processed and summarised using both python and Tableau. For a library and record of all files and changes, visit: [https://github.com/megnicd/programming-for-big-data\\_CA04](https://github.com/megnicd/programming-for-big-data_CA04)

The folder contains the original logfile, the python file used to process the logfile, a python test suite, a .csv file written by the python processing file, a packaged Tableau workbook used to create some summary visuals and an exported crosstab from Tableau. This document will also summarise any findings, and some basic summary information can also be obtained by running the python processing file in the command prompt terminal.

## The Basics

There are 5255 lines of data in the GitHub log file, all of which contain data relating to 422 commits to a single project. The commits were made by 10 different authors between July 13<sup>th</sup>, 2015 and November 27<sup>th</sup>, 2015.

### Figure 1: Commits per author

#### Commits per author











Author		
Thomas		191
Jimmy		152
Vincent		26
/OU=Domain Control Vali..		24
ajon0002		9
Freddie		7
Alan		5
Nicky		5
Dave		2
murari.krishnan		1

Figure 1 is a count of the number of commits made by each author throughout the project. Thomas and Jimmy were by far the most productive in terms of commits. Combined, they made 343 commits - over 80% of the project's total commits.

### Figure 2: Comments per author

#### Comments per author











Author		
Thomas		234
Jimmy		154
Vincent		80
/OU=Domain Control Vali..		24
ajon0002		24
Freddie		14
Alan		8
Nicky		14
Dave		2
murari.krishnan		1

Figure 2 displays the numbers of comments per author. Unsurprisingly, Thomas and Jimmy made the most comments. If we assume each comment is a change; Thomas made 1.2 changes per commit, Jimmy made 1 change per commit, while Vincent made 3 changes per commit.

It could be argued that although Vincent did not make the most commits, his commits were the most meaningful.

**Figure 3: Commits per week**

Commits per week (2015)

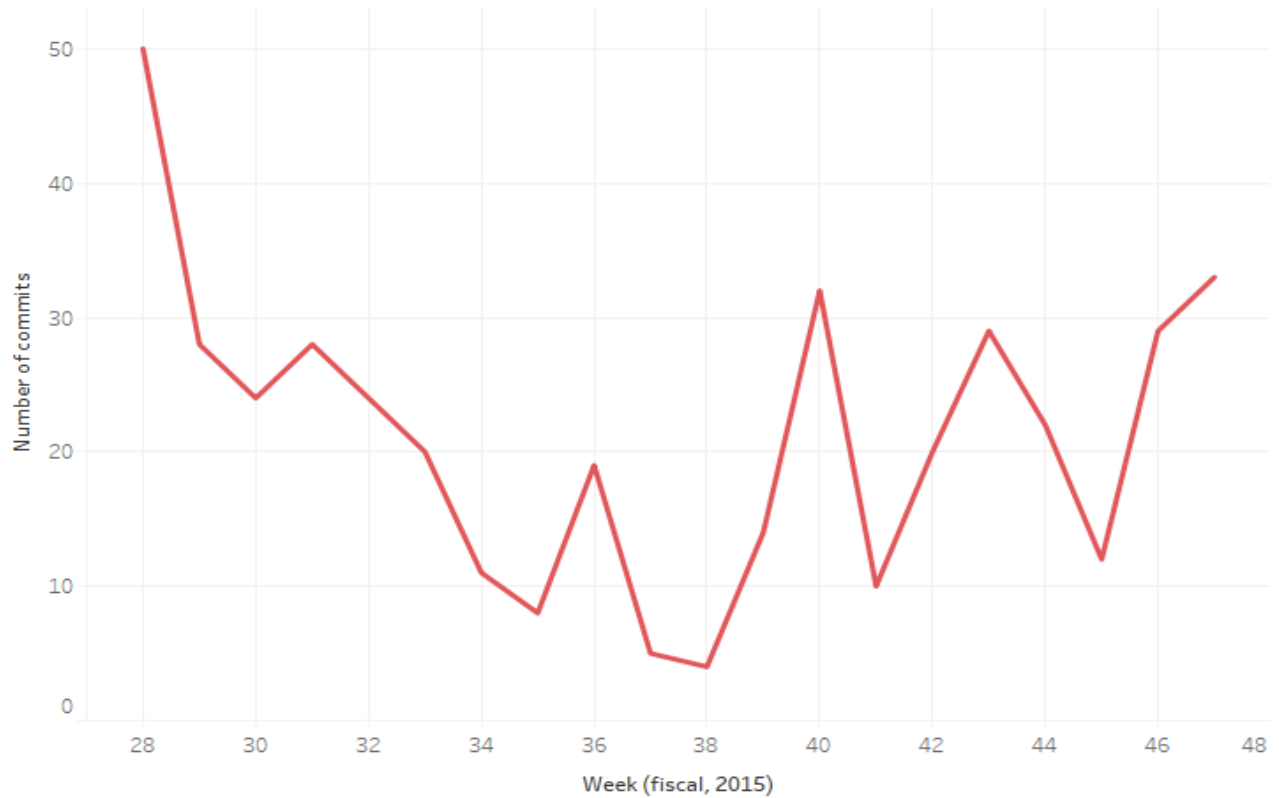


Figure 3 displays the number of commits per week. The largest number of commits were made during week one of the project. The number of commits gradually decreased over time, with a small peak in week 36, followed by two larger peaks in weeks 40 and 43. The number of commits also seemed to be on the increase again toward the end date. In Tableau, you can filter this visual by author.

**Note:** Week 28 (fiscal, 2015) is week 1 of the project. The tooltip in Tableau displays this explicitly.

**Figure 4: Commits per day**

Commits per day (2015)

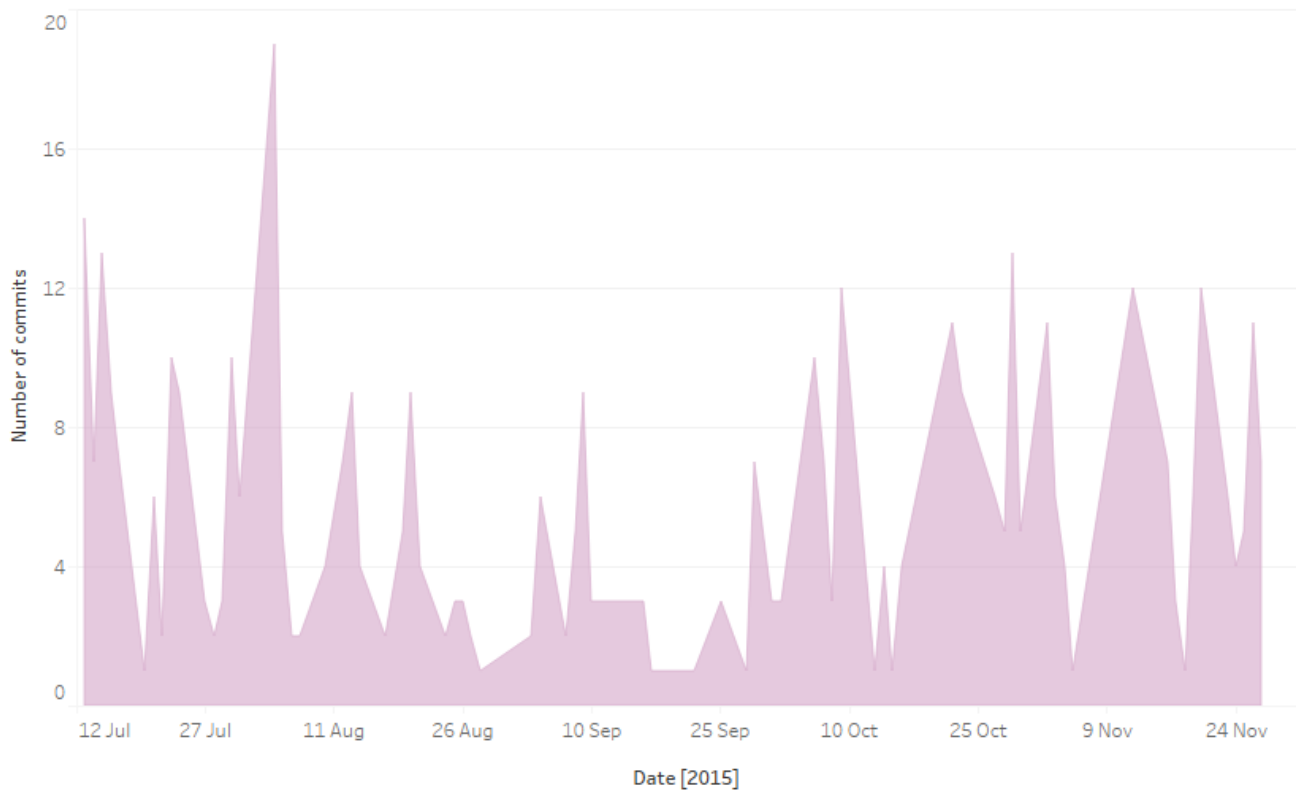


Figure 4 details the number of commits per day. August 4th had the most commits, followed by October 29th. We may have expected a date in week 1 (fiscal week 28) to have the most commits, but the large number of commits in week 1 is due to an above average number of commits over multiple days, rather than a peak in a single day. In Tableau, you can filter this visual by author.

An exported crosstab of this visual also tell us when each author made their very first commit. Alan, Jimmy and Thomas were committing changes to the project from day 1, while /OU=Domain and Nicky joined on days 3 and 4. Dave began contributing later in July, while Freddie, Marari joined in late August. Vincent and Ajon were the last to join in late October and early November.

This not only gives us extra information about the project, but it may help to explain the seemingly low productivity of certain developer. For example, Vincent's 26 commits seem impressive given the additional context that this visual and the data beneath it provide.