Department of Computer Science

Community Analysis Report

Pandas

Author:
Koushik Ganesan

Supervisor:
Prof. Jed Brown

November 22, 2019

1 Introduction

Pandas is an open source, BSD licensed, library in Python offering high-performance and efficient data structures for data analysis and manipulation. It aims to become the most flexible and powerful data analysis tool of any language.

Some important features it offers :

- Easy handling of NaNs in floating point or non-floating point data
- Size mutability; restructuring of DataFrames made intuitive
- Offers the ability to write/load data from a vareity of file formats: CSV, Excel, databases, and the HDF5 format
- Hierarchical axis indexing makes working with higher dimensional data easy to understand with a simpler lower dimenisonal representation

It has been highly optimized with certain parts of code written in C or Cython. It's dependencies include NumPy, python-dateutil and pytz. It also has a code coverage of 93%.

2 History and funding

Development of Pandas began in 2008 at AQR Capital Managememnt by Wes McKinney, a quant researcher. He wanted to develop pandas because he found python user-friendly but was troubled by the lack of its ability to process large amounts of data. He pushed for it to become open source in 2009 because he felt that a large community of python users would benefit from it. In 2015 it became a a fiscally sponsored project of NumFOCUS, a non-profit charity based in the US. Donations made towards the project are exempt from taxes, though a small fees is taken by NumFOCUS towards legal matters. It also has institutional partners whch include : Two Sigma, Anaconda and University of Paris-Saclay. They employee people whose primary job is to work on the development of pandas.

3 Governance

The decisions are made the Core team and the BDFL (Wes McKinney). The core team consists of people who made significant contributions to pandas, although there is no specific metric on which they are scored. The BDFL has the ability to override the decisions made by the Core team but this power is rarely exercised. Every major decision requires atleast 80% of the core team (including the BDFL) to participate and atleast 2/3rd of the members to vote for it inorder to pass. Members with a conflict of interest in a particular issue may participate in Core Team discussions on that issue, but must recuse themselves from voting on the issue. If the BDFL has recused his/herself for a particular decision, they will appoint a substitute BDFL for that decision. More on governance can be found at https://github.com/pandas-dev/pandas-governance.

4 Statistics

Till now there has been 108 version releases with more than 6M downloads, primarily through conda. It has had about 100 coontributors in its lifetime with more than 10 commits and an overall of 1600 conotributors. It has an active and decentralized community with about 350 commits in the last month (Oct-Nov 2019). Most of the discussions happen through gitihub's Issues tab, pandas-dev mailing list and the Gitter Channel.

5 Code of Conduct

Contributors are expected to abide by its "Contributor Code of Conoduct". Any issues, related to this, reported is addressed by a working community of members. This code of conduct is adapted from Contributor Covenant, version 1.3.0, available at http://contributor-covenant.org/version/1/3/0/, and the Swift Code of Conduct.