Data to Insight briefing note: Open Source Software for Data Analysis in Councils

June 2021

Overview

Data to Insight, a local government initiative, supports a longstanding network of local authority children's services data professionals with data tools, learning opportunities and spaces for collaboration.

As well as supporting existing data tools, we lead projects to introduce colleagues to new ways of working. In the contemporary analytical landscape, this often means using open source software alongside traditional tools. We know this poses interesting challenges not only to the analysts learning new skills, but also to the network and software administrators responsible for providing employees with the capabilities they need for their jobs, and for maintaining a secure, safe, and manageable local IT infrastructure.

This briefing note outlines the importance of open source software to contemporary data analysis work. Our goal is to reassure administrators that tools can be used safely in public sector contexts, and to demonstrate to service leaders the value they offer when used by children's services analysts as part of a data-literate organisation.

What is open source software?

Open source software is software with source code that anyone can inspect, modify or enhance.

This does not mean that specific software is entirely unmanaged; the most popular open source projects have created markets for managed versions and packages with a range of support and licensing options to help organisations manage both the effort of working with software, and the risk of installing software.

There exist many good and comprehensive guides to open source software. A useful overview can be found here.

What are the benefits?

The open source analytical tools we recommend represent new capabilities for some analysts, and for others they represent better ways of doing things. While many councils have invested in data visualisation software like Power BI or Tableau, few have invested in data analysis software. These open source tools can be used to address that gap, and generate insights which, especially in the high cost / high risk environment of children's social care, can be the difference between safeguarding a child or not; between effectively procuring care placements or overspending.

We have seen from Ofsted inspection reports that data quality, and the quality of associated analysis, is becoming a key factor in managing a successful children's services department. If we want to help councils make better use of data, we cannot afford to ignore the analytical component of the work which these software support. Those councils who use Python or R do so because their use of the tools generates value far beyond that which was possible using traditional spreasheet and visualisation tools, and because the tools are cheap to start using and easy to scale.

What software do we recommend for local authority data analysts?

Just as with any other software, new tools or versions of tools appear to replace older tools. With our partners across local government we currently see Python and R as our most valuable analytical prospects.

Python is a beginner-friendly programming language with strong support from analyst programmers. We use Python as part of Anaconda, a well-maintained distribution of Python that contains all the programs and packages that our learners will need to get started. We use Jupyter notebooks, a browser-based Python interface included in the

Anaconda distribution, to conduct advanced analytics on varied datasets and present findings in varied and engaging ways. The packages we mainly use when working with Python are:

SciPy, which contains some of the basic data structuring and visualisation tools used by other packages
pandas for manipulating tabular data
scikit-learn and statsmodels for regression, clustering and time series analysis
seaborn for easy to use data visualisations
Natural Language Toolkit — NLTK for text analysis

There are other packages available for creating more advanced data visualisations, connecting to Excel spreadsheets or to SQL databases, and more. Many of these will already be installed with Anaconda. For those that are not, we use the Conda package manager that comes with Anaconda, which is one of the easiest and most reliable ways to manage packages in Python. The package lists Conda defaults to using are maintained by Anaconda and NumFOCUS. Python, Conda and all the Python packages described above are free for commercial use and open source.

R is a free software environment for statistical computing and graphics. Where previously many local authorities have struggled to justify the cost of purchasing dedicated statistical software for their analytical teams, R offers the capabilities most useful for statistical analysis without a purchase cost. It is also now possible to work with R code within SQL Server, making it highly adaptable to existing analytical environments.

How do I know it's safe?

As we have seen during the NHS ransomware attack in May 2017 and countless other similar cases, any software is only as safe as its users and its administrators, and this is equally true of open source software.

The distributions we recommend are professional products. While the code is open source, distributors control their releases. The key risks are the same ones inherent in tools like Excel which every employee can use – namely, that a user might choose to run code without understanding it, or an administrator might neglect a critical update.

Through the charity NumFOCUS, the Python data analysis packages we recommend are maintained and kept secure. The importance of these packages in live systems can be seen in the extensive list of companies who financially support their maintenance. Continuum Analytics, the company that maintains the Anaconda distribution, helped found NumFOCUS and continues to support their projects.

We are also in good company. Not only do global companies like Google, Facebook, and Instagram rely on Python as part of their analytics ecosystem, these tools are also well used in the UK public sector. The Government Digital Service has often recommended Python and Jupyter "for the rapid development and prototyping of solutions to data science problems". A growing number of councils are also now moving forward with some form of advanced analytics, and often using open source software as part of their approach. Southend also offers a useful case study.

How long will it take?

This will depend on your local approach. The basic Anaconda installation is extremely easy to perform and requires no elevated permissions. We understand that councils differ in the assessments they perform prior to approving new software, and councils preferring to virtualise software will incur the usual packaging overheads for the IT service.

What about database permissions?

Simply installing a piece of software does not grant the user any new database permissions or new access rights. It is standard practice to grant data analysts the ability to read data from reporting databases directly, and to set up automations to facilitate their work. This can be done without compromising the live databases in any way.

Many organisations also provide their analysts with reporting databases or warehouses which they can modify themselves. Whether or not your organisation does this will depend on the split of roles, responsibilities and skills between your analysis teams, your software teams, and your database administrators.

Feedback / contact

To contribute any feedback or requests, please email alistair.herbert@eastsussex.gov.uk or join our Slack.