SESUG Paper 23-2024

Benefits, Challenges, and Opportunities with SAS® and Open-source Software (OSS) Integration

Kirk Paul Lafler, sasNerd
Ryan Paul Lafler, Premier Analytics Consulting, LLC
Joshua J. Cook, M.S., ACRP-PM, CCRC
Stephen B. Sloan, MBA, M.S., Dawson DR (Data Research)
Anna T.K. Wade, Premier Analytics Consulting, LLC

ABSTRACT

With all the power and functionality that SAS® has to offer, it now provides users with exciting integration paths to the world of open-source software (OSS). This presentation will create a dialog to engage audience members about strategies, techniques, and integration pathways for integrating the world of SAS software with open-source initiatives and technologies. Technology in the 21st century is experiencing a paradigm shift where organizations around the globe are demanding that all software live, play, and thrive in the same sand box together. We will also discuss the challenges facing the user community as they grapple with methods and approaches to best integrate SAS with open-source software, compatibility and vulnerability issues, security limitations, intellectual property issues, warranty issues, and inconsistent developer practices. Plan to join us for an exciting presentation about the benefits, challenges, and opportunities confronting us all with the integration of SAS software with the incredible opportunities coming from the open-source community, including cloud computing architecture, open standards, and the collaborative nature of community.

Keywords: sas, open-source, open source, oss, python, r, sql, anaconda, linux, firefox, quarto, oss benefits, oss challenges, oss opportunities, oss integration, oss example, data scientist, it

INTRODUCTION

Open-source software (OSS) has become increasingly popular among enthusiasts, particularly in the information technology (IT) and data science (DS) industries. This paper introduces the reader to the distinct software types, the virtues that OSS and its vibrant community of experts provide, OSS examples, the benefits, challenges, and opportunities associated with OSS integration, and the efforts for OSS standardization. OSS benefits include source code transparency, flexibility, agility, identification of security issues, speed of fixing bugs, licensure, and maintenance fee cost-savings.

The adoption and integration of open-source software solutions into businesses' proprietary products represents a turning point from prior decades. Moreso, this integration comes with challenges including version control, inconsistent stability updates, sparse and variable documentation, and potential software vulnerabilities from unstable releases. As a balancing act, individuals, businesses, and organizations must weigh the potential challenges of open-source software with its democratizing philosophy: offering accessible, customizable, integrated, free-to-use, and community-tested products for all. This paper presents several popular, widely available open-source software solutions for commercial, private, and academic uses.

SOFTWARE TYPES

Software is created using source code which tells a program or application how to function. For this paper, two distinct software types (or categories) will be illustrated: 1) Proprietary (or commercial) software and 2) Open-source software. A major decision confronting a software developer is whether the source code related to the software release will be made publicly available on web hosting services, such as GitHub, for anyone to inspect, modify, enhance, and share – referred to as open-source, versus software where the developer maintains exclusive control over the source code thereby preventing the public availability to inspect, modify, enhance, and share it – referred to as closed source or proprietary software.

So, which type of software is more common? A large majority of apps, games, and other popular software are classified as closed source or proprietary. However, there are a growing number of open-source alternatives for users to choose from. For example, a popular open-source alternative to Microsoft Office is LibreOffice. An open-source alternative to Microsoft Windows is the Linux operating system. Another popular open-source alternative to Google's or Bing's web browser software is the Mozilla Firefox web browser.

ECOSYSTEM OF LEADING-EDGE TECHNOLOGIES

A growing number of organizations are launching initiatives to integrate and use leading-edge commercial technologies and open-source software, applications, and tools for the purpose of engineering methodologies and toolkits to meet the needs of organizations worldwide. In an announcement from McKinsey & Company (September 26, 2023), the launch of a technology-driven ecosystem boasts commercial and open-source software, applications, and tools coexisting together - like a biological community of organisms interacting together within a single physical environment. This initiative represents an ecosystem of integrated technologies working together for the purpose of gaining the greatest value from their technology investments.

EXAMPLES OF OPEN-SOURCE SOFTWARE (OSS) APPS AND TOOLS

An alphabetical list of popular examples of OSS applications and tools is presented in table 1.

Open-source Software (OSS)	Description
Anaconda	Anaconda is a distribution of the Python and R programming languages for scientific computing and data science packages that aim to simplify package deployment and management under Windows, Linux, and macOS operating systems.
Apache Hadoop	Apache Hadoop is a collection of open-source software utilities for data science projects that facilitates the integration of a network of computers to solve problems involving massive amounts of data and computing capacity.
Apache HTTP Server	The Apache HTTP Server is a free and open-source cross-platform web server software, released under the terms of Apache License 2.0. Apache HTTP Server allows users to deploy their websites on the Internet.
Apache Mahout	Apache Mahout is an environment for building scalable machine learning algorithms.
Data Version Control	Data Version Control (DVC) is a popular open-source tool used to version data, annotate it with metadata, track changes to data, and collaborate with others on data science projects.
Firefox	Mozilla Firefox is free and open-source software, released under the terms of the Mozilla Public License which means you may use, copy, and distribute Firefox to others.
GDAL	The Geospatial Data Abstraction Library is an open-source, cross-platform spatial data management program. Capable of importing, accessing, manipulating, and exporting geospatial data in several raster file and vector file formats, GDAL is widely accessible through APIs for Python, R, PHP, and Java.
GIMP	GNU Image Manipulation Program (GIMP) is freely distributed software for image composition, image retouching, and image restoration.
jQuery	jQuery is free, open-source software using the permissive MIT license consisting of a JavaScript library that simplifies the creation and navigation of web applications.
Knime	Knime is a free, open-source data analytics, reporting, integration, and machine learning platform for data scientists.
LibreOffice	LibreOffice is powerful and free open-source office software suite for word processing, spreadsheets, presentations, graphics, flowcharts, databases, and formula editing.

Open-source Software (OSS)	Description
Linux	Linux is released under an open-source license which lets anyone use, run, modify, and redistribute the source code and sell copies of their modified source code if they do so under the same license. The source code for Linux is under copyright by its many individual authors and licensed under the General Public License Version 2 (GPLv2) license.
Matplotlib	Matplotlib is an open-source software graph plotting library for the Python programming language.
NumPy	NumPy is a library for the Python programming language and used for scientific computing tasks.
Orange	Orange is an open-source data science toolkit for developing, visualizing, and testing data mining workflows.
PostgreSQL	PostgreSQL is an open-source object-relational database management system, supporting both relational (SQL) and non-relational (JSON) and offering advanced SQL functions, including foreign keys, subqueries, and triggers. PostgreSQL is commonly used as the main data warehouse to support internet-scale web, mobile, and geospatial applications. Importantly, PostgreSQL runs on most machines and operating systems.
Project Jupyter	Project Jupyter provides an environment to develop open-source software applications and tools to perform data cleaning, statistical computation, data visualization, and create predictive machine learning models.
Python	Python is a powerful open-source programming language that is available under a free software license. It supports object-oriented and structured programming along with other programming paradigms. Developed by Guido van Rossum in the late 1980s, Python is designed to be an “easy to read language” with numerous third-party modules to interact with other languages; extensive support libraries such as web service tools; text processing; string operations; internet protocols; a powerful scripting language; an extensive user community; and many other features.
Quarto	Quarto is an open-source scientific and technical publishing system, frequently referred to as “the next generation of R markdown.” Quarto allows users to author using both Jupyter notebooks or with plain text markdown in their editor of choice, with dynamic content in Python, R, Julia, and Observable. The Quarto system allows for reproducible, production-quality articles, presentations, dashboards, websites, blogs, and books in a wide range of formats (ex: HTML, PDF, and more).
QGIS	QGIS is an open-source spatiotemporal software suite capable of viewing, modifying, manipulating, and exporting geospatial data through a graphical user interface. It is the open-source equivalent to the proprietary ArcGIS geospatial software suite. QGIS supports tiling systems; merging and mosaicking raster files; importing and exporting vector shapefiles; delivering customizable data visualizations; interactive mapmaking; spatiotemporal summary statistics; and other GIS features.
R	R is a powerful open-source programming language and is used for statistical computing, graphics, and data analysis. Available under a free software license, R runs on all important platforms and is used by statisticians, data miners and thousands of major corporations and institutions worldwide. Developed by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand, their initial version of R was released in 1995 with a stable beta version in 2000. R boasts an extensive array of packages including data wrangling; data analysis; plotting; graphing; reporting; statistics; an extensive user community; and many other features.
Shiny	Shiny is an open-source R package and Python library that provides an elegant and powerful web framework for building web applications, including interactive data exploration tools, dashboards, and full-blown applications. Shiny helps you turn your analyses into interactive web applications without requiring HTML, CSS, or JavaScript knowledge.

Open-source Software (OSS)	Description
SQL	Structured Query Language (SQL) is a relational database language that is used in managing and programming relational database management systems (RDBMS). SQL is specifically useful in handling structured data and comprises many types of statements including a data query language (DQL), a data definition language (DDL), a data control language (DCL), and a data manipulation language (DML). MySQL and Microsoft SQL Server Express are open-source database applications with a codebase that is free to view, download, modify, distribute, and reuse.
tidyverse	The tidyverse is a collection of R packages that allow for data reading, cleaning, wrangling, and graphing using a standardized syntax on top of the core R programming language. The tidyverse is among the most frequently utilized packages in the R ecosystem, particularly for data analysts and data scientists. The "tidy" syntax introduced by these packages serves as a key design principle for many other R packages, including packages like tidymodels.
VLC Media Player	VideoLAN Client (VLC) Media Player is a free open-source multi-platform media player for desktop operating systems and mobile platforms such as Android, iOS, and iPadOS that plays multimedia files including DVDs, Audio CDs, and many video formats and streaming content.
VNC	Virtual Network Computing (VNC) is an open-source application that provides screen sharing services. VNC is secure over the Internet with all end-to-end connections encrypted and remote computers are protected by a password or by a system login credentials.

Table 1. Open-source Software Descriptions, continued

EXAMPLE OF OSS INTEGRATION – A SYMBIOTIC SYSTEM

The OSS applications and tools described above are not typically used in isolated environments, but instead in symbiotic relationships with each other, working together toward a common goal. An example of OSS integration for data science is provided in Figure 1 below.

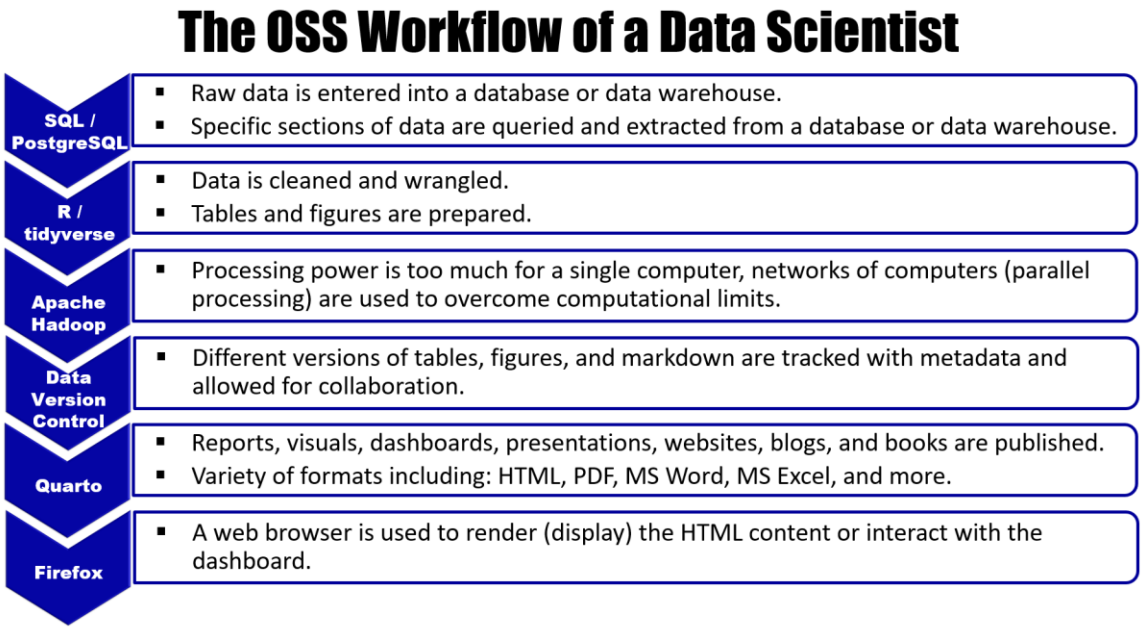

Figure 1. The OSS workflow of a data scientist.

OSS INTEGRATION – BENEFITS, CHALLENGES, AND OPPORTUNITIES

OSS integration promotes free access to inspect, modify, enhance, and share source code. The redistribution of software is not only permitted but encouraged to sustain innovation. According to a recent study by Gartner, open-source tools provide flexibility and cost-effectiveness for data integration tasks and projects such as connectivity, data routing, and transformation.

IS OPEN-SOURCE SOFTWARE (OSS) BUG-FREE?

The short answer to this question is no. But no matter the type of software – commercial or open-source – bugs (or code flaws that affect quality, performance, and security) are inevitable. The primary difference between both types of software is who is ultimately responsible for fixing the bugs. For commercial software, vendors and their professional developers are responsible for fixing bugs; while for open-source software, the legions of users – like you and me – are responsible for fixing bugs.

OSS STANDARDIZATION EFFORTS

There are many efforts across the OSS landscape that are focused on standardization, specifically regarding version control and quality assurance. Within the R ecosystem, the pharmaverse is a network of companies aimed at curating open-source R packages that are standardized for clinical reporting in the pharmaceutical industry. Other groups that are working toward standardization in R include PHUSE Working Groups, R Consortium Working Groups, R Validation Hub, and TransCelerate. Importantly, on September 27th, 2023, the R Submissions Working Group successfully completed the first publicly available submission package to the U.S. Food & Drug Administration (FDA) that included a Shiny component.

In the Python ecosystem, one can observe similar endeavors to those in the R ecosystem, like the creation of centralized repositories and working groups dedicated to enhancing the utility and reliability of Python packages. Projects such as PyPA (Python Packaging Authority) work to improve the packaging landscape and develop standards for Python package distribution. The Python Software Foundation also promotes and protects Python through community-driven efforts.

In the realm of data science and machine learning, groups such as NumFOCUS support and organize standards around popular Python packages like NumPy, pandas, and scikit-learn, ensuring they meet the high standards required for academic and industrial research. Additionally, industry collaborations, like the Consortium for Python Data API Standards, aim to create a uniform API for array/tensor-like data structures across various libraries, improving interoperability and user experience.

These efforts aim to directly address many of the challenges associated with OSS.

CONCLUSION

Open-source software (OSS) has become increasingly popular among enthusiasts particularly in the IT and DS industries. This paper introduced the reader to the distinct software types, the virtues that OSS and its vibrant community of experts provide, OSS examples, OSS integration examples, the benefits, challenges, and opportunities associated with OSS integration, and the efforts for OSS standardization. Specific benefits of OSS include source code transparency, flexibility, agility, identification of security issues, speed of fixing bugs, licensure, and maintenance fee cost-savings.

Open-source software is available for a variety of applications. From entire programming languages, geospatial software suites (GIS programs), libraries, packages, and modules, data warehouses, data lakes, and structured databases. Furthermore, open-source software, applications, and tools offer transparent, cost-effective, and customizable software solutions that can be integrated into any organization's professional workflow.

REFERENCES

Gartner Research (26-August-2019). "What Innovation Leaders Must Know About Open-Source Software."

Lafler, Kirk Paul, Ryan Paul Lafler, Joshua J. Cook, Stephen B. Sloan, and Anna T. K. Wade (2024). "Benefits, Challenges, and Opportunities with SAS and Open-source Software (OSS) Integration," Proceedings of the 2024 Pharmaceutical SAS Users Group (PharmaSUG) Conference.

Lafler, Kirk Paul and Ryan Paul Lafler (2023). "Benefits, Challenges, and Opportunities with SAS and Open-source Software (OSS) Integration," Proceedings of the 2023 Western Users of SAS Software (WUSS) Conference.

Lafler, Kirk Paul (2019). *PROC SQL: Beyond the Basics Using SAS, Third Edition*, SAS Institute Inc., Cary, NC, USA.

McKinsey & Company (September 26, 2023). "McKinsey launches an open-source ecosystem for digital and AI projects."

R Consortium (September 27, 2023). "Shiny App Successfully Reviewed by FDA CDER Staff (Pilot 2 Announcement 2)."

ACKNOWLEDGMENTS

The authors thank the SESUG 2024 Conference Committee, particularly the Careers, Training, and Education Section Chairs, Jonathan Duggins and John Betz, for accepting our paper; the SESUG 2024 Academic Chair, Jim Blum, and the Operation Chair, Lucia Alexander, for organizing and supporting a great "in-person" conference event; SAS Institute Inc. for providing SAS users with wonderful software; and SAS users everywhere for being the nicest people anywhere!

TRADEMARKS

AUTHOR CONTACT INFORMATION

Kirk Paul Lafler is a consultant, developer, programmer, educator, and data scientist; and teaches SAS Programming and Data Management in the Statistics Department at San Diego State University. Kirk also provides project-based consulting and programming services to client organizations; and teaches "virtual" and "live" SAS, SQL, Python, Database Management Systems (DBMS) technologies (e.g., Oracle, SQL-Server, Teradata, MySQL, MongoDB, PostgreSQL, AWS), Excel, R, and cloud-based technologies. Currently, Kirk serves as the Western Users of SAS Software (WUSS) Executive Committee (EC) Open-Source Advocate and Coordinator. Kirk is the author of several books including the popular PROC SQL: Beyond the Basics Using SAS, Third Edition (SAS Press. 2019); an Invited speaker, educator, keynote, and leader; and the recipient of 29 "Best" contributed paper, hands-on workshop (HOW), and poster awards.

Ryan Paul Lafler is the Founder, C.E.O., Chief Data Scientist, and Lead Consultant at Premier Analytics Consulting, LCC, a consulting and contracting business that specializes in Big Data Science products and services for clients. Ryan also serves as an Adjunct Professor at San Diego State University for the Master of Science Big Data Analytics (BDA) Program and the Department of Mathematics and Statistics. He received his Master of Science in Big Data Analytics from San Diego State University after defending and publishing his Thesis and graduated with Honors in May 2023. Ryan's specialties include programming in Python, R, SAS, JavaScript, and SQL for data science, machine learning engineering, deep learning integration, statistical analysis, spatiotemporal analysis, data visualization, interactive dashboard development, and database structuring purposes.

Joshua J. Cook, M.S. DS, M.S. CRM, ACRP-PM, CCRC, is a dedicated professional with a robust background in bench to bedside research, aiming for a career as a physician-scientist. He has completed two concurrent master's degrees, led in the clinical research and data science industries, published and presented extensively, and holds certifications by ACRP as a Project Manager and Clinical Research Coordinator. Currently, he serves as a NIH Graduate Research Fellow at the University of South Carolina Big Data Health Science Center and as an Adjunct Professor at the University of West Florida. Joshua is applying to dual doctoral (M.D./Ph.D.) programs with a clear goal to integrate biomedical sciences, clinical research, and data science to enhance evidence-based patient care and research development. He values teaching and mentorship, aspiring to guide others as his mentors did for him.

Stephen B. Sloan has worked in a variety of functional areas including Project Management, Data Management, and Statistical Analysis for Human Resources, Supply Chain, Finance, Marketing, Insurance, Life Sciences, and Manufacturing on behalf of both private and government clients. Stephen has had the good fortune to have worked with many talented people at SAS Institute. Stephen has presented over 100 times at 47 SAS conferences and been published in professional journals. Stephen has a B.A. cum laude with Honor in Mathematics from Brandeis University, M.S. degrees in Mathematics and Computer Science from Northern Illinois University, an MBA from Stern Business School at New York University (1st in his class), and a graduate certificate in Financial Analytics from Stevens Institute.

Anna Wade is a statistician based in San Diego with expertise in statistical analysis, data visualization, and environmental research. She holds a Master's in Statistics with a concentration in biostatistics from San Diego State University, where she graduated with honors. Her diverse experience spans education, consulting, health care, and environmental research, where she combines skills in SAS, R, and Python to tackle complex data challenges. She is currently acting as a Consultant with Premier Analytics Consulting, LLC., and through her work hopes to inspire the next generation of researchers and scientists, all while advocating for ethical practices, equal opportunities, and environmental stewardship.

Comments and suggestions are encouraged and can be sent to:

Kirk Paul Lafler, sasNerd
Consultant, Developer, Programmer, Data Scientist, Educator, and Author
Specializing in SAS® / Python / SQL / Database Management Systems / Excel / R / AWS / Cloud-based Technologies
E-mail: KirkLafler@cs.com
LinkedIn: https://www.linkedin.com/in/KirkPaulLafler/
Twitter: @sasNerd

~ ~ ~ ~ ~ ~ ~

Ryan Paul Lafler, M.Sc.
Premier Analytics Consulting, LLC *and* San Diego State University
Founder, CEO, Chief Data Scientist, Lead Consultant, and Adjunct Faculty
E-mail: rplafler@premier-analytics.com
Website: www.Premier-Analytics.com
LinkedIn: www.LinkedIn.com/in/RyanPaulLafler/

~ ~ ~ ~ ~ ~ ~

Joshua J. Cook
Adjunct Professor, University of West Florida
E-mail: jcook0312@outlook.com
LinkedIn: https://www.linkedin.com/in/joshua-j-cook-934075169/

~ ~ ~ ~ ~ ~ ~

Stephen B. Sloan
CEO, Dawson DR (Data Research)
E-mail: stephen.stephensloan@gmail.com
LinkedIn: https://www.linkedin.com/in/stephen-b-sloan/

~ ~ ~ ~ ~ ~ ~

Anna T. K. Wade, M.Sc. in Statistics
Premier Analytics Consulting, LLC
Statistician, Mathematician, and Consultant
E-mail: annaw21435@gmail.com
Website: https://www.premier-analytics.com/