Paper 143-2023

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

Kirk Paul Lafler, sasNerd
Ryan Paul Lafler, Premier Analytics Consulting, LLC

Abstract

The open-source world is alive, well and growing in popularity. This paper highlights the many benefits found with open source software (OSS) including its flexibility, agility, talent attraction, and the collaborative power of community; the trends show that open-source is ubiquitous penetrating many critical technologies we depend on, where more technology companies recognize the importance of the open-source community leading to more initiatives and sponsorships that support open-source creators; the challenges of open source including compatibility vulnerability issues, security limitations, intellectual property issues, warranty issues, and inconsistent developer practices; and the opportunities coming out of the open source community including cloud architecture, open standards, and the collaborative nature of community.

Introduction

Open-source software (OSS) has increasingly become more popular among enthusiasts particularly in the IT industry. This paper introduces the reader to the distinct software types; the virtues that OSS and its vibrant community of experts provide; OSS examples; and the benefits, challenges, and opportunities associated with OSS integration. OSS benefits include source code transparency, flexibility, agility, identification of security issues, speed of fixing bugs, license, 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 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 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) Applications and Tools

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

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.
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.
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.

Open-source Software (OSS)	Description
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.
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 many other popular 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.
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.
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.

OSS Integration – Benefits, Challenges, and Opportunities

Open-source software 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.

Conclusion

Open-source software (OSS) has increasingly become more popular among enthusiasts particularly in the IT industry. This paper introduced the reader to the distinct software types; the virtues that OSS and its vibrant community of experts provide; OSS examples; and the benefits, challenges, and opportunities associated with OSS integration. Specific benefits of OSS include source code transparency, flexibility, agility, identification of security issues, speed of fixing bugs, license, 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; and many more, 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 (2019). *PROC SQL: Beyond the Basics Using SAS, Third Edition*, SAS Institute Inc., Cary, NC, USA.

Lafler, Kirk Paul (2013). *PROC SQL: Beyond the Basics Using SAS*, Second Edition; SAS Institute Inc., Cary, NC, USA.

Lafler, Kirk Paul (2003). *PROC SQL: Beyond the Basics Using SAS*; SAS Institute Inc., Cary, NC, USA.

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

Acknowledgments

The authors thank the WUSS 2023 Conference Committee, particularly the Open Source Section Chairs, Isaiah Lankham and Matthew Slaughter, for accepting our paper; the WUSS 2023 Academic Chair, Lida Gharibvand, and the Operation Chair, Julie Kilburn, for organizing and supporting a great "live" conference event; SAS Institute Inc. for providing SAS users with wonderful software; and SAS users everywhere for being the nicest people anywhere!

Trademarks

About the Authors

Kirk Paul Lafler is an educator, developer, programmer, consultant, and data analyst; currently working as a lecturer and adjunct professor at San Diego State University and the University of California San Diego Extension; and teaching SAS, SQL, Python, Excel, and cloud-based technology courses to users around the world. Kirk has decades of programming experience and specializes in SAS software, SQL, RDBMS technologies (Oracle, SQL-Server, Teradata, DB2), Python, and other languages and productivity tools. Kirk is the author of the popular PROC SQL: Beyond the Basics Using SAS, Third Edition (SAS Press. 2019) and is actively involved with SAS, SQL, Python, R, ML, and cloud-computing user groups, conferences, and blogs as an Invited speaker, educator, keynote, and leader; and is the recipient of 27 "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.

Comments and suggestions can be sent to:

Kirk Paul Lafler, sasNerd
SAS® / SQL / RDBMS / Python / Cloud-based Consultant, Developer, Programmer, Data Analyst, Educator and Author
E-mail: KirkLafler@cs.com
LinkedIn: https://www.linkedin.com/in/KirkPaulLafler/
Twitter: @sasNerd

~ ~ ~ ~ ~ ~ ~

Ryan Paul Lafler, M.Sc. in Big Data Analytics
Premier Analytics Consulting, LLC
Big Data Scientist, Consultant, and Trainer
E-mail: rplafler@premier-analytics.com
Website: https://www.premier-analytics.com/
LinkedIn: https://www.linkedin.com/in/ryanpaullafler/