PharmaSUG 2024 - Paper SI- 452

Embracing Diversity in Statistical Computing Environments: A Multi-Language Approach

Amit Javkhedkar, Ephicacy Analytics.
Sridhar Vijendra, Ephicacy Analytics.

ABSTRACT

In the transition to Statistical Computing Environments (SCEs), organizations face the challenge of supporting multiple programming languages, including SAS, R, Python, and Julia. Managing these many programming languages, their versions, libraries/packages, and IDEs is challenging.

This paper explores possible ways to standardize and deploy multiple versions of various programming languages within a single platform. It delves into the advantages, limitations, and potential strategies to enhance the experience for the key users in statistical programming teams. Specifically, we examine whether there exists a straightforward solution to this complex task, ultimately aiming to provide insights into optimizing programming and application development in open-source languages for statistical programmers.

INTRODUCTION

In recent years, the statistical programming landscape has been transformed with the transition from traditional personal desktop-based SAS systems to modern web-based platforms and the introduction of R programming languages such as R and Python for analyses and reporting.

As the pharmaceutical sector continues to evolve, organizations are increasingly recognizing the necessity of employing a diverse range of programming languages to effectively address the complex analytical and visualization requirements of clinical trials. This recognition underscores the need to implement systems and processes capable of supporting different technologies, thereby enabling a unified and comprehensive approach to data analysis for submission.

Regulatory bodies such as the FDA and EMA place paramount importance on the ability to audit and trace data analysis processes. This emphasis on auditing and traceability aims to ensure the integrity, reliability, and compliance of data analysis procedures with regulatory standards. In the event of audits or inspections, pharmaceutical companies must be equipped to demonstrate a clear and transparent trail of data manipulation, analysis, and reporting activities.

As we embark on this exploration, our focus extends beyond merely identifying potential solutions for integrating multiple programming languages. We must also address the challenges inherent in achieving robust auditing, traceability and reproducibility functionalities within these systems. By navigating these challenges effectively, we can pave the way for enhanced efficiency, compliance, and innovation in the realm of statistical programming for clinical trials.

VALIDATED SYSTEM

THE ADVANTAGES OF A VALIDATED SYSTEM, SUCH AS SAS, ARE MANIFOLD AND SIGNIFICANT WITHIN THE PHARMACEUTICAL AND CLINICAL RESEARCH INDUSTRIES.

1. REGULATORY ACCEPTANCE: SAS, AS A FULLY VALIDATED SYSTEM, HOLDS ACCEPTANCE FROM GLOBAL REGULATORY AUTHORITIES SUCH AS THE FDA, EMA, AND PMDA. THIS RECOGNITION ENSURES COMPLIANCE WITH STRINGENT REGULATORY STANDARDS AND FACILITATES SMOOTHER APPROVAL PROCESSES FOR CLINICAL TRIALS AND SUBMISSIONS.

2. SIMPLIFIED DEPLOYMENT: DEPLOYING SAS IS STRAIGHTFORWARD AND HASSLE-FREE. WITH THE ENTIRE SAS APPLICATION BEING VALIDATED, ORGANIZATIONS ENCOUNTER MINIMAL CHALLENGES RELATED TO VERSION COMPATIBILITY OR PACKAGE VALIDATION. THIS SIMPLIFIES THE IMPLEMENTATION PROCESS AND REDUCES POTENTIAL DISRUPTIONS TO WORKFLOW.

3. RELIABILITY AND TRUST: THE VALIDATION OF SAS INSTILLS CONFIDENCE IN THE RELIABILITY AND ACCURACY OF ITS OUTPUTS. DATA ANALYSIS PERFORMED USING SAS IS TRUSTED BY REGULATORY AUTHORITIES AND STAKEHOLDERS, PROVIDING ASSURANCE OF DATA INTEGRITY AND QUALITY IN CLINICAL RESEARCH OUTCOMES.

4. CONSISTENCY AND STANDARDIZATION: SAS OFFERS A STANDARDIZED PLATFORM FOR DATA ANALYSIS, ENSURING CONSISTENCY ACROSS ANALYSES AND PROMOTING STANDARDIZATION WITHIN ORGANIZATIONS. THIS CONSISTENCY STREAMLINES PROCESSES, ENHANCES EFFICIENCY, AND FACILITATES EASIER COLLABORATION AMONG TEAM MEMBERS.

5. ACCEPTED ANALYSES: ANALYSES CONDUCTED USING SAS ARE WIDELY ACCEPTED BY REGULATORY AUTHORITIES, ELIMINATING CONCERNS REGARDING THE VALIDITY OR CREDIBILITY OF ANALYTICAL RESULTS. THIS ACCEPTANCE EXPEDITES THE REVIEW AND APPROVAL OF CLINICAL TRIAL SUBMISSIONS, REDUCING TIME-TO-MARKET FOR NEW TREATMENTS OR THERAPIES.

THE INDUSTRY HAS COME TO RELY ON SAS(R) AS IT HAS OVER THE DECADES BECOME THE STANDARD PROGRAMMING LANGUAGE FOR PHARMACEUTICAL AND CLINICAL RESEARCH, RANGING FROM REGULATORY COMPLIANCE AND DUE TO SIMPLIFIED DEPLOYMENT AND REPRODUCIBILITY..

OPEN-SOURCE

THE ADVANTAGES OF OPEN-SOURCE PROGRAMMING LANGUAGES SUCH AS R AND PYTHON ARE MULTIFACETED AND IMPACTFUL, PARTICULARLY IN THE CONTEXT OF COLLABORATIVE DEVELOPMENT AND INNOVATION.

1. CROWD CONTRIBUTION: ONE OF THE MOST SIGNIFICANT ADVANTAGES OF OPEN-SOURCE SYSTEMS IS THE ACTIVE PARTICIPATION AND CONTRIBUTION FROM A DIVERSE COMMUNITY OF DEVELOPERS AND USERS. THIS COLLABORATIVE APPROACH LEADS TO CONTINUOUS IMPROVEMENT AND EVOLUTION OF THE SYSTEM, WITH NEW FUNCTIONALITIES AND FEATURES BEING ADDED REGULARLY. THE COLLECTIVE EXPERTISE AND CREATIVITY OF THE COMMUNITY ENHANCE THE SYSTEM'S CAPABILITIES AND ADDRESS EMERGING NEEDS EFFECTIVELY.

2. RAPID DEVELOPMENT: OPEN-SOURCE SYSTEMS BENEFIT FROM RAPID DEVELOPMENT CYCLES FACILITATED BY THE COLLECTIVE EFFORTS OF CONTRIBUTORS. WITH

DEVELOPERS WORLDWIDE COLLABORATING ON CODE, BUG FIXES, AND FEATURE ENHANCEMENTS, UPDATES AND IMPROVEMENTS ARE ROLLED OUT AT A FASTER PACE COMPARED TO PROPRIETARY SYSTEMS. THIS AGILITY ENSURES THAT THE SYSTEM REMAINS CURRENT AND ADAPTABLE TO CHANGING REQUIREMENTS AND TECHNOLOGIES.

3. FLEXIBILITY AND CUSTOMIZATION: OPEN-SOURCE SYSTEMS OFFER UNPARALLELED FLEXIBILITY AND CUSTOMIZATION OPTIONS. USERS HAVE ACCESS TO THE SOURCE CODE, ALLOWING THEM TO TAILOR THE SYSTEM TO THEIR SPECIFIC NEEDS AND PREFERENCES. THIS LEVEL OF FLEXIBILITY ENABLES ORGANIZATIONS TO CREATE BESPOKE SOLUTIONS THAT ALIGN CLOSELY WITH THEIR UNIQUE WORKFLOWS AND REQUIREMENTS, FOSTERING INNOVATION AND DIFFERENTIATION IN THEIR RESPECTIVE FIELDS.

4. COST-EFFECTIVENESS: OPEN-SOURCE SYSTEMS ARE TYPICALLY AVAILABLE FREE OF CHARGE OR AT A SIGNIFICANTLY LOWER COST COMPARED TO PROPRIETARY SOFTWARE. THIS AFFORDABILITY MAKES THEM ACCESSIBLE TO ORGANIZATIONS OF ALL SIZES, INCLUDING STARTUPS AND NON-PROFIT ORGANIZATIONS, DEMOCRATIZING ACCESS TO POWERFUL TOOLS AND TECHNOLOGIES. MOREOVER, THE ABSENCE OF LICENSING FEES ALLOWS ORGANIZATIONS TO ALLOCATE RESOURCES MORE EFFICIENTLY TOWARDS OTHER CRITICAL AREAS OF THEIR OPERATIONS.

5. TRANSPARENCY AND SECURITY: THE TRANSPARENCY OF OPEN-SOURCE SYSTEMS FOSTERS TRUST AND CONFIDENCE AMONG USERS. WITH ACCESS TO THE SOURCE CODE, USERS CAN REVIEW AND AUDIT THE SYSTEM FOR SECURITY VULNERABILITIES, ENSURING GREATER TRANSPARENCY AND ACCOUNTABILITY. ADDITIONALLY, THE COLLABORATIVE NATURE OF OPEN-SOURCE DEVELOPMENT PROMOTES PROACTIVE IDENTIFICATION AND RESOLUTION OF SECURITY ISSUES, LEADING TO ROBUST AND RESILIENT SOFTWARE SOLUTIONS.

OVERALL, OPEN-SOURCE SYSTEMS OFFER A COMPELLING ARRAY OF ADVANTAGES, INCLUDING CROWD CONTRIBUTION, RAPID DEVELOPMENT, FLEXIBILITY, COST-EFFECTIVENESS, TRANSPARENCY, AND SECURITY. THESE BENEFITS MAKE OPEN-SOURCE SOFTWARE AN ATTRACTIVE CHOICE FOR ORGANIZATIONS SEEKING INNOVATIVE, CUSTOMIZABLE, AND COST-EFFICIENT SOLUTIONS TO MEET THEIR DIVERSE NEEDS AND CHALLENGES.

NEED FOR MULTIPLE LANGUAGE

The need for multiple programming languages in the clinical industry is becoming increasingly evident in today's data-driven landscape. With data being recognized as a valuable asset, the field of data science plays a crucial role in aiding business leaders in making critical decisions. Specifically in the realm of clinical trials, clinical data scientists are tasked with extracting, analyzing, and modeling data to uncover patterns and provide insights into the progress of trials.

Through data analysis, clinical trial costs have seen significant reductions as efficiencies are gained. Data-driven insights enable quicker identification of trial sites and patients, ultimately leading to faster decision-making regarding the effectiveness of treatments. This accelerated decision-making process can potentially shorten the overall duration of clinical trials, benefiting both patients and pharmaceutical companies.

Moreover, the availability of numerous open-source tools has democratized data analysis, making it more accessible and cost-effective for researchers and organizations within the clinical industry. By leveraging a variety of programming languages and tools, clinical data scientists can adapt to diverse data sets and analytical requirements, ultimately enhancing the efficiency and effectiveness of clinical trial processes.

SUBMISSION CHALLENGES WITH OPEN-SOURCE LANGUAGE

Submission challenges arise when using open-source languages such as R due to the dynamic nature of package development and validation processes. While exploratory work benefits from the flexibility and innovation of open-source tools, ensuring compliance and reproducibility in submissions presents unique hurdles.

With the continuous influx of new packages in open-source ecosystems, organizations must navigate the validation process to ensure the reliability and integrity of these tools. Third-party vendors play a crucial role in validating packages but managing multiple validated packages for different analyses within a study can be complex.

Reproducibility is paramount for submission integrity, but it becomes challenging with open-source software like R. Unlike proprietary systems, where software versions and dependencies are controlled, maintaining consistency across different versions of R and its associated packages poses a significant challenge. Without meticulous package management, reproducing results at a later date may prove difficult or impossible.

Furthermore, ensuring auditability is essential for compliance and traceability in submissions. Audit trails provide valuable insights into actions performed, who performed them, and when they occurred. However, the availability of robust audit trails in open-source languages may vary, posing a potential risk for traceability and metadata management.

Regardless of the Integrated Development Environment (IDE) used for program development, maintaining traceability and auditability is critical. Organizations must implement robust processes and tools to track changes, document actions, and ensure compliance throughout the submission lifecycle. By addressing these challenges, organizations can harness the benefits of open-source languages while meeting regulatory requirements and maintaining submission integrity.

STATISTICAL COMPUTING ENVIRONMENT (SCE) – A PARADIGM SHIFT

The Statistical Computing Environment (SCE) can indeed serve as a solution to the challenges mentioned above. SCE platforms are designed to provide a comprehensive environment for statistical programming and data analysis tasks, offering features such as integration with multiple programming languages, advanced visualization tools, and robust auditing and traceability functionalities.

1. **Package Management**: SCEs provide centralized repositories for managing packages, ensuring that only validated and approved versions are used in analyses. Automated dependency tracking and version control mechanisms help maintain consistency across projects and minimize the risk of using incompatible or unvalidated packages.

2. **Reproducibility**: SCEs offer features to reproduce analyses precisely, regardless of changes in package versions or software environments.

3. **Traceability**: SCEs enable the documentation of metadata, including information about data sources, transformations, and analysis steps. By linking input data to analysis outputs and documenting the workflow, SCEs ensure traceability and accountability throughout the research process.

4. **Inbound and Outbound Integration**: SCEs support seamless integration with external systems and data sources through APIs, data connectors, and standardized file formats. This enables researchers to import data from diverse sources and export results for further analysis or reporting.

5. **Standardization**: SCEs enforce standardized folder hierarchies, naming conventions, and analysis structures, promoting consistency and organization across projects. This standardization simplifies collaboration, enhances reproducibility, and facilitates knowledge sharing within research teams.

6. **Global Roles and Permissions**: SCEs allow administrators to define granular access controls and assign roles with specific permissions. This ensures that only authorized users can access

sensitive data or perform critical actions, reducing the risk of unauthorized changes or data breaches.

7. **Collaborative Workflow**: SCEs provide collaborative tools such as shared workspaces, version control systems, and project management dashboards. These tools facilitate communication, task assignment, and progress tracking, enabling efficient collaboration among team members working on complex research projects.
8. **Reliable Auditing**: SCEs maintain comprehensive audit logs that track user actions, changes to code or data, and system configurations. These logs provide a detailed record of all activities, promoting transparency and facilitating compliance with regulatory requirements. Audit logs are tamper-proof and securely stored, providing reliable evidence for compliance audits, investigations, and quality assurance processes.
9. **Cloud Computing**: SCEs leverage cloud computing infrastructure to offer flexible deployment options, scalable computing resources, and high availability. Researchers can access computing resources on-demand, scale up or down as needed, and deploy analyses in distributed environments for improved performance and cost-effectiveness.
10. **Auto-Scalability and Dynamic Storage Expansion**: SCEs automatically scale computing resources and expand storage capacity in response to changing workload demands. This ensures optimal performance, minimizes downtime, and eliminates the need for manual intervention to provision or manage infrastructure resources.

In summary, SCEs provide comprehensive solutions to the challenges of using open-source languages in clinical research. By addressing package management, reproducibility, auditability, traceability, integration, standardization, collaboration, access control, and scalability, SCEs empower researchers to conduct rigorous and compliant research while maximizing efficiency and productivity.

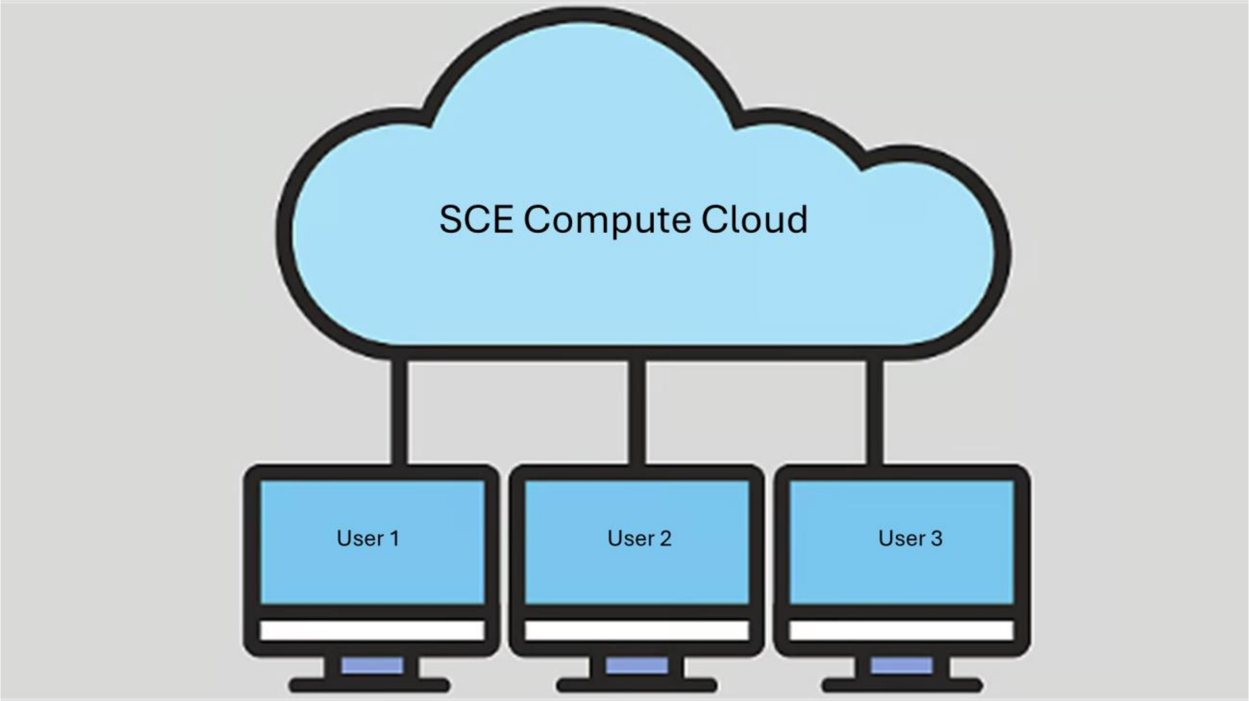

Figure 1 – SCE – A cloud based single environment encompassing all the Statistical Programming needs.

MEETING BUSINESS NEEDS AND REGULATORY NEEDS

In meeting both business and regulatory needs, modern Statistical Computing Environments (SCEs) offer essential functionalities that cater to the diverse requirements of data scientists and regulatory compliance standards. By integrating various Interactive Development Environments (IDEs) such as RStudio, VS Code, and Jupyter Notebook, SCEs empower programmers and data scientists to work seamlessly in their preferred editor, fostering productivity and flexibility. Moreover, SCEs provide robust role-based access control, ensuring that users only have access to specific areas of analysis while maintaining full auditability and traceability—a critical aspect in adhering to stringent regulations. This segregation of roles minimizes interference between programming and end-user access, safeguarding the integrity of analyses. Additionally, SCEs support the creation of interactive web applications, leveraging packages like R Shiny for complex dataset analysis. The ability to configure Python and R runtimes within SCEs enables users to develop and deploy applications utilizing popular frameworks like Python Streamlit and RShiny, enhancing collaboration and decision-making capabilities. Furthermore, SCEs facilitate reproducibility through versioning and traceability features, enabling users to reproduce analyses accurately even amidst the continuous evolution of packages and functions in open-source languages like R. By offering these functionalities, SCEs empower organizations to meet both business objectives and regulatory requirements effectively in today's dynamic landscape.

CONCLUSION

In conclusion, modern Statistical Computing Environments (SCEs) emerge as indispensable tools in meeting the complex demands of the clinical research industry. By seamlessly integrating various Interactive Development Environments (IDEs) and providing robust role-based access controls, SCEs facilitate efficient collaboration among data scientists while ensuring compliance with stringent regulatory standards. The ability to develop interactive web applications and support multiple programming languages enhances analytical capabilities and decision-making processes, driving innovation and productivity. Moreover, SCEs' emphasis on reproducibility through versioning and traceability features instills confidence in the integrity of analyses, even in the face of evolving open-source languages and packages. As organizations navigate the challenges of data analysis and regulatory compliance, SCEs serve as invaluable assets, enabling them to achieve their business objectives while upholding the highest standards of quality, transparency, and accountability in clinical research endeavors.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Sridhar Vijendra
Ephicacy Analytics
sridhar.vijendra@ephicacy.com

Amit Javkhedkar
Ephicacy Analytics
amit.javkhedkar@ephicacy.com