

Bridging Platforms: Integrating RStudio POSIT and SAS Grid in the same environment

Sebastia Barcelo, Cytel, Geneva, Switzerland

Malte Stein, Cytel, Berlin, Germany

ABSTRACT

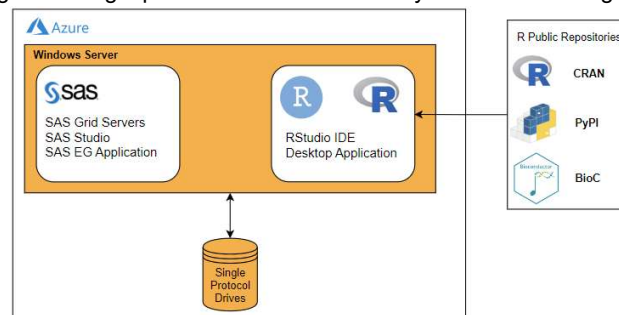
Integrating RStudio POSIT into the same environment where we have SAS Grid posed several challenges due to disparate platforms - SAS Grid installed on Windows server and POSIT installed on Linux server.

This presentation outlines the hurdles encountered during the integration process and the innovative solutions devised to overcome them. From addressing compatibility issues to streamlining access to the same drives from both platforms, we navigated through complexities to ensure seamless integration.

Through collaborative efforts and strategic planning, we successfully bridged the gap between Windows and Linux environments, enabling efficient utilization of RStudio POSIT and SAS Grid accessing the same data. Join us to explore our journey, learn from our experiences, and gain insights into effectively integrating diverse technologies within a unified environment.

INTRODUCTION

The integration of advanced data analytics tools is crucial for enhancing efficiency and flexibility in modern computational environments. In our initial infrastructure, we relied on a SAS Grid system, a robust solution for managing large-scale, distributed data processing. The SAS Grid was installed on a Windows-based server and used by a team that also accessed data through RStudio Desktop (using packages from public repositories), which was similarly installed on the Windows server. This setup allowed users to work within the same environment, accessing shared resources on a designated single protocol drive that was only accessible through Windows applications.



While functional, this infrastructure had limitations, particularly in terms of flexibility and scalability. The growing demand for open-source tools such as RStudio POSIT, with its suite of capabilities, including Workbench, Package Manager, and Connect, presented an opportunity to enhance our data analysis capabilities. However, POSIT tools require a Linux-based server for installation and operation, creating a disparity between the two environments. Our goal was to integrate POSIT Workbench, which provides an interactive data science environment, Package Manager, which centralizes package management and distribution, and Connect, a publishing platform for sharing analytical content, into the existing infrastructure while ensuring seamless access to the same data resources.

This integration effort posed several challenges, particularly due to the disparate operating systems and the need to enable shared access to resources across both Windows and Linux platforms. To address this, we embarked on a journey to build a unified environment, focusing on resolving compatibility issues, streamlining data access, and ensuring efficient use of both SAS Grid and POSIT tools. In this paper, we explore the hurdles we faced and the solutions that allowed us to bridge the gap between these two platforms, creating a cohesive and efficient data analysis ecosystem.

IMPLEMENTING THE NEW INFRASTRUCTURE

The implementation of the new infrastructure required careful planning to ensure smooth integration between the

existing SAS Grid environment and the newly added POSIT tools. This section outlines the key areas addressed during the implementation process, with a focus on internal vs. external access, data storage, and other integration challenges.

INTERNAL VS EXTERNAL ACCESS

The implementation of internal and external access for POSIT Connect presented both advantages and challenges that needed to be carefully balanced to meet organizational requirements for security, performance, and compliance, while keeping in mind operational costs and user management complexities.

Pros

- **Security:** One of the primary advantages of segregating internal and external access is enhanced security. By having distinct environments, it becomes easier to apply different security policies and user authentication mechanisms tailored to the needs of internal or external users. For instance, internal users may benefit from single sign-on (SSO) with Active Directory (AD), while external users could require multi-factor authentication (MFA), ensuring robust security protocols for each group.
- **Data Segregation:** Separating internal data from external access significantly reduces the risk of exposing sensitive information to unintended audiences. This segmentation ensures that confidential or proprietary data remains within the organization's secure network, while only non-sensitive content is shared with external stakeholders.
- **Performance:** The infrastructure can allocate dedicated resources for internal users, ensuring that external traffic does not impact the performance of internal workflows. This separation mitigates the risk of external users consuming excessive bandwidth or computational power, which could otherwise slow down operations for internal teams.
- **Compliance:** Keeping internal and external content separate helps organizations meet compliance requirements, especially in scenarios where external-facing content must adhere to different legal or regulatory standards. For example, there might be restrictions on certain data packages or libraries available to external users due to licensing agreements or data protection laws, while internal users might not be subject to these limitations.

Cons

- **Increased Maintenance and Higher Costs:** Maintaining separate environments for internal and external users increases the complexity of the infrastructure, leading to higher operational costs. This includes the need for additional servers, maintenance work, and security monitoring, which would otherwise be simpler if both internal and external users shared a common platform.
- **Complex User Management:** With separate environments, user authentication and authorization mechanisms become more complex to manage. Administrators must track user access across multiple systems and ensure that appropriate roles and permissions are applied consistently across internal and external domains.
- **Integration Challenges:** Publishing content from POSIT Workbench directly to external facing Connect instances adds complexity to workflows. Since external content cannot be published directly from the internal instance of Workbench (and we decided to not have a Workbench instance on the external server), users must first publish internally and then replicate the content for external access. This introduces additional steps and processes into the workflow, increasing the time and effort required to make content available to external users.

Balancing these pros and cons was crucial in our approach to enabling internal and external access to POSIT Connect, ensuring security, performance, and compliance while addressing the operational challenges posed by maintaining separate environments.

DATA STORAGE

Ensuring seamless access to shared data between the SAS Grid (Windows) and the POSIT (Linux) environment was a critical aspect of the integration process. To address this, we implemented dual-protocol drives that support both SMB (Server Message Block) for Windows and NFS (Network File System) for Linux. This solution allowed for efficient cross-platform data sharing, resolving compatibility issues that initially prevented POSIT's access to the same datasets used by the SAS Grid.

The dual-protocol drives enabled both platforms to read and write to the same data repository without the need for additional data transfer mechanisms or manual synchronization. This integration streamlined workflows, allowing users of both SAS Grid and POSIT tools to work with the same datasets in real-time, improving efficiency and collaboration.

However, migrating to dual-protocol drives introduced some challenges, particularly regarding folder permissions. Mapping the differences between Active Directory (AD) permissions on Windows and Linux-based file permissions at the subfolder level proved complex. While AD handles permissions in a certain hierarchy, Linux file systems rely on a different structure, which required detailed configuration to ensure that permissions were applied consistently across both platforms.

Due to these challenges, the migration was not immediate. For a transitional period, we continue using single-protocol drives to avoid disruptions to ongoing workflows. The plan is to decommission the single-protocol drives after fully resolving the permissions issues and completing the migration to the dual-protocol system. This will provide a unified storage solution that supports cross-platform access without compromising data integrity or security, ensuring that both SAS Grid and POSIT can seamlessly access shared data.

OTHER INTEGRATION CHALLENGES

In addition to data storage and access, several technical challenges arose during the integration of the POSIT tools.

Workbench Access

During the integration of POSIT Workbench, we encountered instability in the client_init process, which disrupted access to the platform. The instability primarily affected users attempting to connect through their browsers. After investigation, we identified that browser cache buildup, particularly in Chrome, was contributing to the unstable connections.

To address this, we implemented an automatic Chrome cache cleaning process, which helped stabilize connections to POSIT Workbench. By clearing the cache periodically, users were able to access POSIT Workbench platform, and performance became more consistent across different user sessions.

RStudio – Package Manager

The integration between RStudio Desktop and the POSIT Package Manager has been largely seamless, with no significant compatibility issues between Windows (RStudio Desktop) and Linux (Package Manager). This has allowed users to install and manage R and Python packages without major disruptions.

However, we did encounter cache-related challenges during package installation in RStudio Desktop. Some packages failed to install properly due to cache conflicts, requiring additional troubleshooting to be resolved. Adjusting cache settings for each volume ensured smooth package management moving forward.

RStudio/POSIT – SAS Grid

One of the more significant limitations in the integration is the inability to submit SAS jobs directly from POSIT Workbench. As it stands, SAS jobs can only be submitted through RStudio Desktop. This constraint has posed a challenge for certain workflows, particularly when attempting to build Shiny applications that generate and execute SAS code. Since SAS jobs cannot be run directly from POSIT Workbench, users must rely on RStudio Desktop to bridge this gap, limiting the flexibility of POSIT Workbench in executing cross-platform SAS workflows.

POSIT – Git

We also encountered an issue related to triplicated sessions when interacting with GIT via the dual-protocol drive. This anomaly, currently under investigation by POSIT, creates redundant GIT sessions, complicating version control management. Once resolved, this issue will streamline the integration with GIT, allowing for more efficient collaboration and version tracking for all the R projects.

To facilitate publishing applications for external access without directly using POSIT Workbench on the external server, we implemented a git-backed process. This process allows applications and content to be published to POSIT Connect directly from a GIT repository. When new commits are pushed to the repository, POSIT Connect automatically deploys the updated content. This solution not only simplifies the deployment process but also ensures that external-facing applications remain up to date without manual intervention from the internal Workbench instance.

These integration challenges, while complex, were addressed with tailored solutions that maintained the integrity of the overall infrastructure while improving usability and functionality across platforms.

CONCLUSION

Integrating POSIT tools into the existing SAS Grid environment required overcoming a series of technical challenges, primarily due to the disparities between the Windows-based SAS Grid and the Linux-based POSIT infrastructure. Through strategic planning and the implementation of tailored solutions, we were able to bridge these gaps and create a unified, efficient environment that leverages the strengths of both platforms.

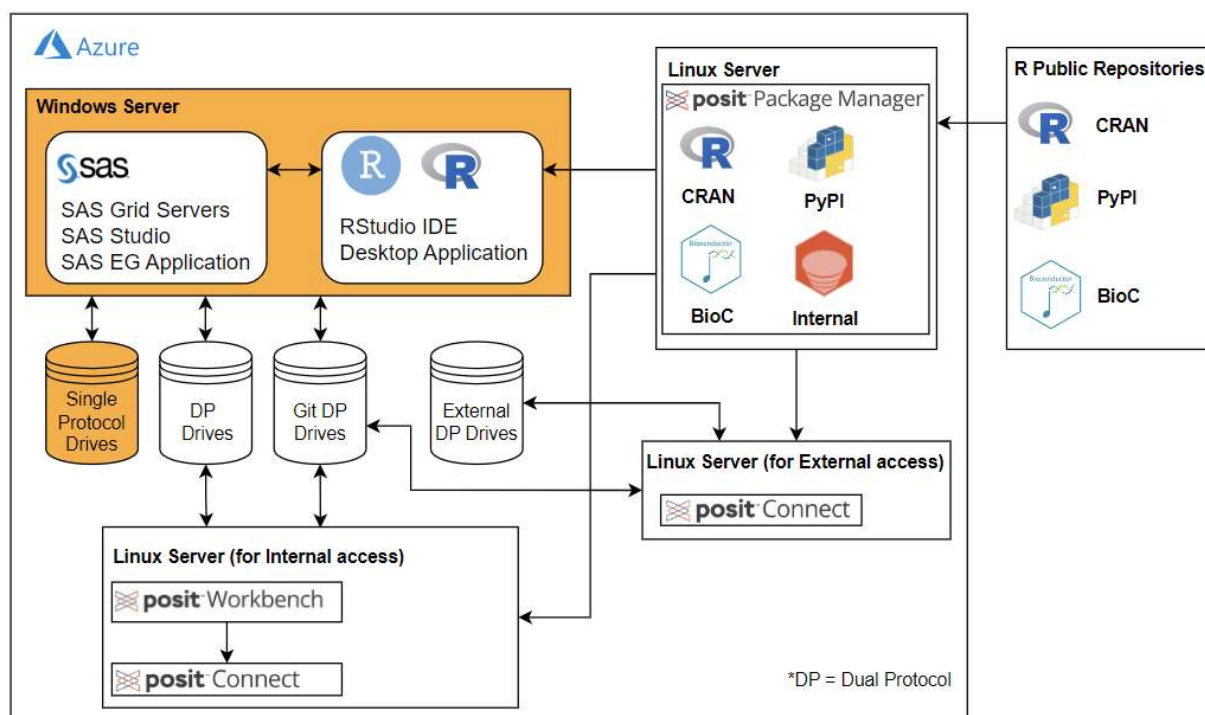
One of the most significant improvements was the introduction of dual-protocol drives, enabling seamless cross-platform data access for both SAS and POSIT users. Although challenges arose around permissions mapping between Active Directory (Windows) and Linux file systems, the move to dual-protocol drives has allowed for efficient, real-time collaboration on shared datasets. While we are still working through the final stages of migration, this solution provides a solid foundation for decommissioning single-protocol drives and streamlining data access across both environments.

Further, the integration of internal and external access to POSIT Connect presented both advantages and operational challenges. By segregating internal and external content, we strengthened security and compliance while ensuring performance remained optimal for both user groups. However, the increased complexity of user management and the higher costs associated with maintaining separate environments necessitated a more complex workflow, especially for publishing content externally.

Other notable integration challenges included stabilizing POSIT Workbench access, ensuring reliable communication between RStudio Desktop and POSIT Package Manager. Each of these challenges were met with practical solutions, such as automated cache clearing for improved Workbench stability, troubleshooting package installation issues, and using GIT-backed deployment to simplify publishing for external users.

Despite these challenges, we have successfully created an integrated system that allows both Windows-based SAS Grid and Linux-based POSIT tools to function cohesively within a shared infrastructure. This integration not only improves data access and collaboration but also provides a scalable foundation for future growth, including the potential for deeper integration with GIT and enhancements to cross-platform workflows.

The following diagram illustrates the final infrastructure:



In conclusion, this integration project demonstrates that, while complex, it is possible to successfully bring together disparate systems like — Windows-based SAS and Linux-based POSIT— by addressing compatibility issues, ensuring data consistency, and streamlining access. The insights gained from our experience offer valuable lessons for any organization looking to integrate diverse technologies within a unified, cross-platform environment.

CONTACT INFORMATION

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

Co-Author Name: Sebastia Barcelo

Company: Cytel

Address: 2nd Floor, Block H, International Center Cointrin, 1215 Geneva 15, Switzerland

Email: Sebastia.Barcelo@Cytel.com

Website: Cytel.com

Author Name: Malte Stein

Company: Cytel

Address: Potsdamer Straße 58, 10785 Berlin, Germany

Email: Malte.Stein@Cytel.com

Website: Cytel.com

Brand and product names are trademarks of their respective companies.