Paper DH13

Data Pipeline – Improving Cycle Time and Shifting Data Closer to Analytics

Jayasree Ivaturi, MaxisIT, Frisco, TX, USA

Sarika Karnik, MaxisIT, Edison, NJ, USA

ABSTRACT

Clinical trials DM and Analytics processes are facing challenges due to the increasing data diversity and sources. Process of sourcing data, managing quality, review, standardization, and statistical programming for regulatory submission involves multiple siloed systems and extensive manual interventions; risking synergies, data integrity and traceability in critical last mile.

Our fully automated end-to-end data pipeline offers timely data access to consumers across the data-to-outcome analysis process with quality, integrity, and end-to-end traceability from source to analysis. These data pipelines bring data from various vendors at a set frequency; verify the quality and adherence of incoming data to standards; and offer oversight of data review observations between the DM and vendors until resolution. Finally, the refreshed data is channeled through the data pipeline to the statistical programmer on demand for regulatory analytics. This approach streamlines data to analytics process, ensuring efficient and accurate delivery of data with integrity and traceability.

INTRODUCTION

This paper focuses on highlighting the current challenges faced by statistical programmers and how our data pipeline improves cycle time and to deliver timely access to quality data for analytics and reporting teams.

MaxisIT's CTR platform has bridged the siloed steps into more streamlined, automated workflows, which form contextual data pipelines of quality data and that can be triggered by clinical data analytics and reporting teams. This feature delivers the power in the hands of A&R teams and overcomes dependencies. Moreover, this feature reduces manual work and addresses the risks associated with isolated processes.

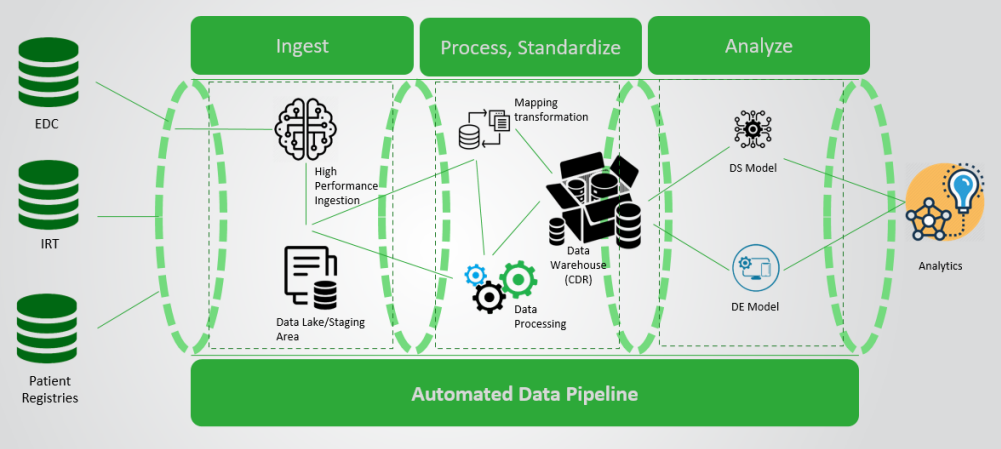
A key focus of this feature is that it explores the transformative impact of a fully automated data pipeline infused with **Continuous Integration/Continuous Deployment (CI/CD)** concepts to meet the unique needs of data scientists and statistical programmers. Companies wish to streamline their data and analytics processes to make them more efficient, provide high quality data faster, while ensuring a stable and reliable operations in general [1] CI/CD framework improves code source control, build automation tools and test code base automatically [2] It significantly impacts the last-minute data requirements crucial for regulatory submissions.

The key concepts leveraged CI/CD concept include:

Agile Data Access with CI/CD	Last-Minute Precision	Impact on Regulatory Submissions
• Introducing agility into the traditionally rigid data management processes. This ensures rapid and reliable data access for data scientists and statistical programmers, facilitating quick iterations and adjustments as needed	• Maintains a focus on precision. Continuous monitoring, validation, and quality checks guarantee not only swift data delivery but also accuracy, even in the eleventh hour.	• With data scientists and statistical programmers empowered by timely, accurate, and rapidly accessible data, the last-minute phase becomes an opportunity for strategic decision-making, reducing the risk of errors and ensuring a smoother path to regulatory approval.

COMPONENTS & PROCESS

Every element within the data pipeline, spanning from raw data to derived insights, is precisely designed to uphold precision while maintaining integrity and traceability.



DATA INGESTION, PROCESSING & STANDARDIZATION

Our approach involves integrating diverse data sources to establish a cohesive dataset, facilitating a comprehensive comprehension of clinical research from various dimensions. This is achieved through an AI-enabled, efficient ingestion process. The system is adept at automatically conducting source validation checks tailored to the specific type of source, ensuring data reliability.

To augment the seamless integration and exchange of data, we implemented data interoperability standards, further enhancing the interoperability of clinical information, which is crucial to the integrity of the dataset are the pivotal stages of data cleaning and preprocessing, where we address missing values and standardize formats. These steps are paramount in refining the dataset, laying a robust foundation for subsequent analysis and interpretation in the dynamic field of clinical data science and statistical programming.

DATA QUALITY MANAGEMENT AND GOVERNANCE

The integrated approach to data quality management and governance involves implementing mechanisms for verifying the quality of incoming data and conducting processes for cleansing and standardizing data to maintain consistency and adherence to predefined standards. Concurrently, the establishment of a robust metadata management system

facilitates the documentation and tracking of the data's origin, transformations, and overall quality. Ensuring end-to-end traceability from the source data to the final analysis addresses regulatory requirements and maintains data integrity throughout the entire lifecycle. This comprehensive strategy aims to not only enhance data quality but also establish governance practices that contribute to operational efficiency and regulatory compliance in the context of clinical trial data management and analytics. “Three Core Principles for Regulatory Compliance & Implementation Success:” the ART model- Accuracy, Reproducibility and Traceability. Accuracy is the correctness measure of program libraries used to generate results from an SCE (Statistical Computing Environment) application. StatOPS dictates accuracy is proven, before SCE is functional. Accuracy can be assured by documenting its Program Development Lifecycle processes, testing methodologies and results, indicating required processes were followed. The best way to demonstrate accuracy is to fully test the program's functionality against requirements. [4]

DATA ANALYTICS

After the standardized data is transmitted to our central data repository, we proceed to construct either Data Science models or Data Engineering models. These models play a pivotal role in generating analytics that are utilized to scrutinize the data and extract meaningful insights. The refreshed standardized or as-is data subsequently traverses the data pipeline, reaching statistical programmers who leverage it for regulatory analytics.

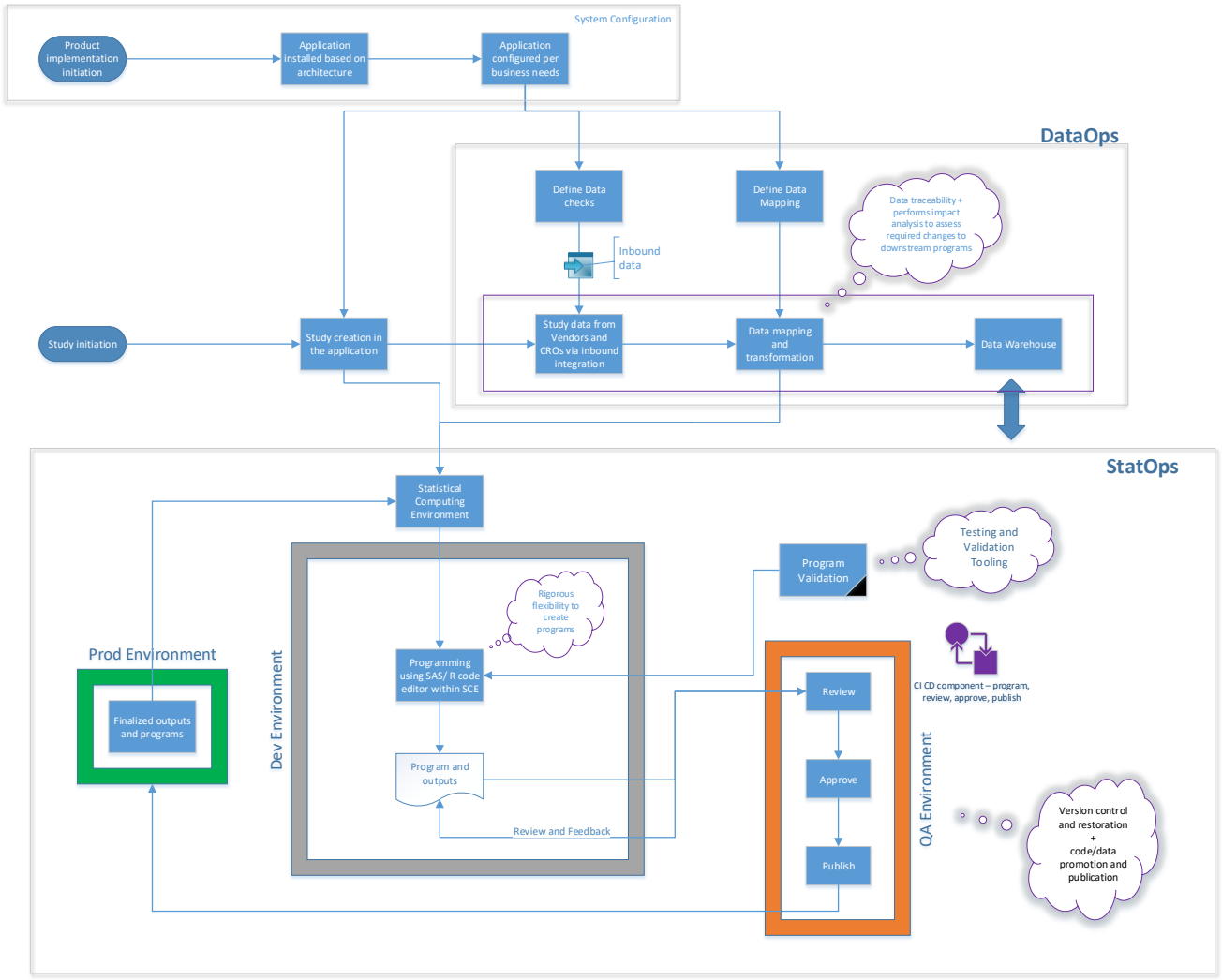
Fully configurable, and automated data pipelines within MaxisIT’s statistical computing environment represent a significant step forward in clinical trials data analytics’ cycle time optimization. It goes beyond a mere technological upgrade by strategically addressing industry challenges to enable clinical trials to harness diverse data sources while upholding rigorous standards of integrity and traceability; and adds computing power within one unified computing environment, while catering to GxP and non-GxP scenarios in on-demand manner. [5]

Maxis IT SCE’s StatOPS process represents a new generation of platform and statistical operations practices focused on productizing the delivery of submission-ready data and analyses, as well as the programs, scripts, algorithms, and processes that generate them. Some of the principles it is built on include:

- Embrace cultural philosophies, best practices and right tools that will increase biopharma organizations’ ability to deliver as well as manage regulatory standard statistical programming and exploratory analytics code development using Agile best-practices.
- Promote ‘rigorous flexibility’ allowing for rich, full-featured and language-agnostic statistical programming development environments tailored to the individual programmer while adhering to organizational, department, and regulatory compliance requirements [5]. Programmers can bring their preferred IDE (Integrated Development Environment) or programming language of choice (e.g., SAS, R, Python etc.). The platform provides bi-directional synchronization with the users' local network drive enabling BYOIDE (Bring Your Own IDE).
- Enable regulatory compliance activities, generating and managing Work Instructions (WI) and documentation required for submission as well as SOPs’ compliance.
- Leverages ‘as-code’ principles to:
    - Standardize statistical computing platform configurations.
    - Improve security and auditability.
    - Harmonize change management.
    - Reduce validation burden and speed deployment timelines.
    - Increase QC (Quality Control) and validation test coverage via automation.
    - Simplify integration with cloud-native applications.
    - Scale system performance
- Promote principles of FAIR data (Findable, Addressable, Interoperable, Reusable) via ‘Subscriber Management’ services which harmonizes provisioning of clinical data pipelines with existing security infrastructure (i.e., user/role file system permissions, directory services, ACLs, etc.)

- Reduce technical debt and improves code reuse of statistical programs and other analytic assets across studies via 'Program Guide' services that build publication libraries of metadata-driven code descriptions containing intended use, methods, pipelines employed, etc. (like git read.me).

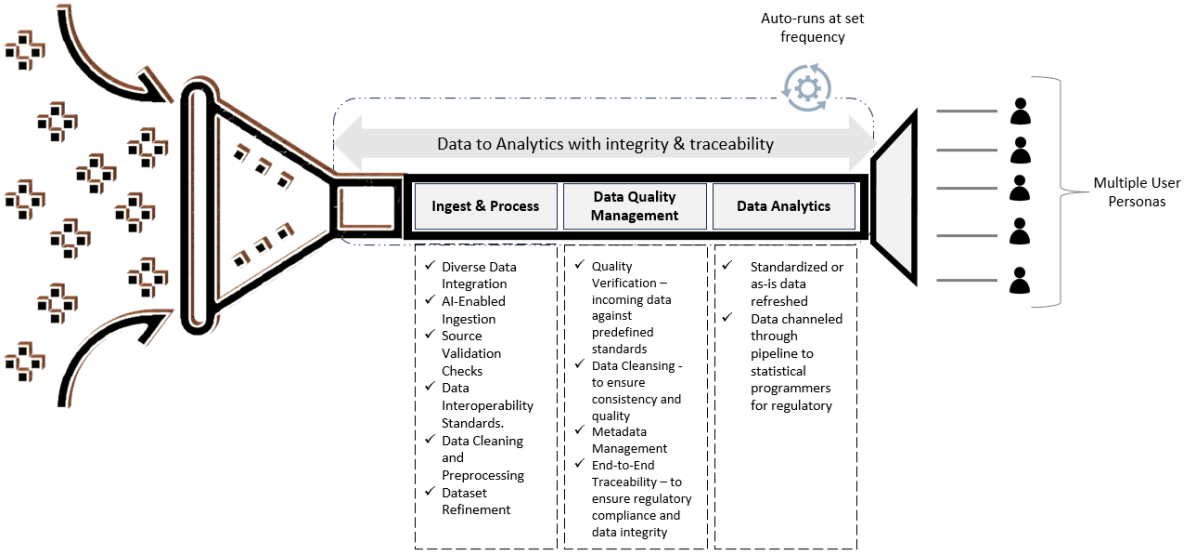
MaxisIT SCE



Exploratory statistical analysis is paramount in opening new avenues of research, preparing for confirmative statistics, providing supportive information, identifying possible undesired effects of therapies and to finding sub-populations, where a therapy is efficacious.[3] It is an innovatively redesigned SCE for the clinical trials regulatory standard as well as exploratory data science programming needs that drives business objectives with policy-based data and resources consumption and successfully operationalizes statistical computing practices in tech-enabled infrastructure by ensuring governance and cycle-time optimization throughout the statistical computing lifecycle.

AUTOMATED DATA PIPELINE

By integrating CI/CD concepts into clinical trial data management workflows[1], we can achieve more efficient collaboration between data handling teams and data scientists as well as statistical programmers, ensuring the reproducibility of analyses, and maintaining the highest standards of data quality and security.



The underlying framework must place a strong emphasis on Data Observability.

DATA OBSERVABILITY [6]

It is indicative of an organization's ability to comprehensively assess the well-being of data within their systems. It serves to diminish the occurrence and impact of data downtime (periods marked by partial, erroneous, missing, inaccurate data) by actively monitoring and alerting teams to potential incidents that might otherwise remain undetected.

Like software observability, data observability is built up on distinct pillars:

a. Volume: Has all the data arrived as expected? Are there any anomalies, duplicates, or missing data?
b. Freshness: Is the data most up to date and most recent?
c. Distribution: Is the data within acceptable ranges and is it within the expected format?
d. Schema: What is the schema currently used and are there any intentional changes?
e. Lineage: What are the upstream and downstream dependencies connected with a specific data asset? Who relies on this data for decision making and where is it stored?

MaxisIT implements data observability through its data management platform which brings together all the above components.

CHALLENGES WITH RESOLUTION

Addressing the challenges below is crucial for the successful implementation of a fully automated end-to-end data pipeline, which aims to streamline the clinical trials Data Management and Analytics processes.

DATA DIVERSITY AND SOURCES

Managing and integrating data from diverse sources, such as different vendors, can be complex and challenging. This diversity may lead to inconsistencies, data quality issues, and difficulties in standardization. With digital data transfer agreements' adherence and API-based data integrations can overcome this.

MANUAL INTERVENTIONS

Extensive manual interventions in the data management and analytics processes can introduce errors, delays, and increase the likelihood of oversight. Manual processes may compromise data integrity, traceability, and overall

efficiency, particularly in the critical last mile of the process. Leveraging automated data quality checks and AI-enabled anomaly detection can help detect issues faster.

DATA QUALITY & ADHERENCE

Verifying the quality and adherence of incoming data to standards is a crucial but challenging aspect of the data management process. Inaccurate or non-compliant data may compromise the validity of analytical results and, consequently, the reliability of regulatory submissions. Our comprehensive strategy aims to enhance data quality and establish governance practices that contribute to operational efficiency and regulatory compliance in the context of the respective working groups.

EFFICIENCY AND ACCURACY IN DATA DELIVERY

Ensuring the efficient and accurate delivery of data to statistical programmers for regulatory analytics is a key concern. Delays or inaccuracies in data delivery may impact the timelines for regulatory submissions and overall trial efficiency. [1] Our automated data pipeline infused with CI/CD ensures end users are getting timely, accurate, and rapidly accessible data with accuracy and reduces the risk of errors and ensuring a smoother path to regulatory approval.

END-TO-END DATA TRACEABILITY

Achieving end-to-end traceability from source to analysis is a critical requirement but can be difficult to ensure. Lack of traceability may hinder the ability to audit and validate the entire data and analytics process, impacting regulatory compliance [4]. Our data pipeline is built to have the traceability starting from the data inception to analytics with a clear data lineage at any given step in the data pipeline for end users ease of review.

CONCLUSION

The increasing diversity and sources of data in these processes often lead to siloed systems, manual interventions, and potential risks to synergies, data integrity, and traceability in the critical last mile. By offering timely data access with a focus on quality, integrity, and end-to-end traceability from source to analysis, the proposed data pipeline efficiently streamlines the entire data-to-outcome analysis process. This approach not only enhances the speed of data access for clinical data scientists and statistical programmers but also ensures accurate delivery with integrity and traceability, thereby mitigating potential risks and improving overall efficiency in regulatory analytics for clinical trials.

REFERENCES

1. “DataOps - Towards a Definition,” Ereth, Julian. (2018).
2. Manasi Vartak (March 10, 2022) What is MLOps? DataOps? And Why do They Matter? https://devops.com/what-is-mlops-dataops-and-why-do-they-matter/
3. Gaus et al., Clin Exp Pharmacol 2015, 5:4 “Interpretation of Statistical Significance - Exploratory Versus Confirmative Testing in Clinical Trials, Epidemiological Studies, Meta-Analyses and Toxicological Screening (Using Ginkgo biloba as an Example)” http://dx.doi.org/10.4172/2161-1459.1000182
4. “Modernization of Statistical Analytics Solutions” Transcelerate Biopharma Inc. https://www.transceleratebiopharmainc.com/assets/modernization-of-statistical-analytics-solutions/
5. Bynens et al., (2021). Statistical Computing Environment [White Paper]. Phuse US Connect. https://www.lexjansen.com/phuse-us/2021/hw/PAP_HoW04.pdf
6. “DataOps Explained: How to Not Screw it up” https://www.montecarlodata.com/blog-what-is-dataops/

CONTACT INFORMATION

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

Author Name: Jayasree Ivaturi & Sarika Karnik

Company: MaxisIT, Email: info@maxisit.com

Address: 510 Thornall Street, Suite 180, Edison, NJ 08837

Work Phone: +1 732-494-2005, ext. 135

Website: https://www.maxisit.com/