Mainframe Data Wrangling: Preparing Your Data for Use in Machine Learning Models

Joshua Powell
Pittsburgh, PA
joshua.powell@broadcom.com

Broadcom

Abstract

Mainframe computing continues to drive the global economy, with forty-five of the world's top fifty banks [1] handling critical transaction data through the IBM Z mainframe platform. While recent research highlights the importance of mainframe modernization rather than replacement [2], enterprises struggle to effectively utilize mainframe data for automation and optimization due to data-driven and communication-driven failures [3]. This challenge creates a significant gap between available mainframe capabilities and realized business value [4].

To address this gap, we conducted semi-structured interviews with eighteen participants across three roles: mainframe subject matter experts (SME) [n=6], mainframe individual contributor end-users [n=9], and mainframe people manager end-users [n=3]. The study, conducted between [dates], investigated two research questions: [RQ1] What are the primary use cases in which mainframe network and network security data impacts mean-time-to-resolution (MTTR) in top fifty banks? And [RQ2] What mainframe data sources and methods do end-users employ to resolve these network and network security issues?

Analysis revealed that 91% of participants [5] encountered data quality or completeness issues that impeded network problem resolution. This tutorial demonstrates how end-users can leverage exploratory data analysis techniques [6], mainframe data APIs [7], and open source data science tools [8] to prepare data for advanced analytics and machine learning applications. The presented methodology aims to reduce MTTR by addressing identified data-driven and communication-driven failure points [9], with specific focus on network security use cases [10-13].

Keywords: Data Engineering, Data Wrangling, API Usability, API Onboarding, Mainframe, Large-scale Computing, Machine Learning

1 Tutorial Aims and Objectives

Data wrangling enables organizations to transform raw mainframe data into actionable insights that can reduce mean-time-to-resolution (MTTR) in critical banking operations [5]. This tutorial addresses the challenges identified through our research with eighteen mainframe professionals across three roles: SMEs, individual contributors, and people managers.

1.1 Stakeholder Communication & Understanding

Our research revealed that teams often begin building solutions before fully understanding stakeholder requirements or establishing effective communication channels [3]. This tutorial presents two user experience research methods specifically designed to:

- Clarify stakeholder goals in network security contexts
- Improve collaboration between technical and business teams

Powell, J. I. Broadcom Mainframe Software. (2024, November 6). Mainframe data wrangling: Preparing your mainframe data for use in machine learning models. [Conference Tutorial]. Guide Share Europe (GSE) UK Conference 2024, - United Kingdom.

- Address the communication-driven failures identified in our research

1.2 Data Quality

With 91% of studied mainframe professionals reporting data quality challenges, establishing high-quality data is crucial for reliable business decision-making and problem-solving [7]. This criticality increases when augmenting operations with artificial intelligence. The tutorial provides:

- Tools and techniques for improving mainframe network data quality
- Prepare: Methods for preparing data
- Explore: Approaches for maintaining data integrity during transformation
- Visualize: Techniques for meaningful visualization

The tutorial's primary objective is to address the data-driven and communication-driven failures identified in our research by:

- Introducing open-source tools for mainframe data wrangling
- Demonstrating REST API usage for mainframe data exploration [8]
- Teaching visualization techniques for analysis and machine learning preparation [9]
- Promoting data quality governance aligned with banking industry requirements

Through these objectives, participants will learn practical approaches to reduce MTTR in network security contexts while maintaining data integrity and stakeholder alignment.

2 Intended Audience and Required Background

This tutorial is designed for technical professionals working with IBM z/OS mainframe environments [1], particularly those involved in network operations and problem resolution. The content is specifically relevant for:

2.1 Primary Audience

- Mainframe Subject Matter Experts (SMEs)
- Network Security Specialists
- System Programmers
- Enterprise Architects
- Data Scientists and Data Engineers
- Application Developers
- DevOps Engineers

2.2 Secondary Audience

- IT Operations Managers
- Technical Project Managers
- Quality Assurance Engineers
- Business Analysts working with mainframe data

2.3 Required Background

- Basic understanding of mainframe architecture and z/OS concepts
- Familiarity with programming concepts and REST APIs
- Basic knowledge of Python programming language
- Understanding of data analysis fundamentals

2.4 Recommended Experience

- 1+ years working with mainframe systems
- Basic exposure to network security concepts
- Familiarity with command-line interfaces
- Basic understanding of data quality principles

No prior experience with machine learning or advanced data analysis is required, though basic statistical knowledge will be helpful.

Note: While the tutorial focuses on IBM z/OS environments [1], the data wrangling principles and methodologies presented are applicable to other large-scale computing environments.

3 Relevance

This tutorial addresses critical challenges in mainframe environments, particularly within banking and financial services where 45 of the top 50 banks rely on IBM Z platforms [1]. The tutorial's relevance is established through three key factors:

1. Growing Demand for Advanced Analytics
   (a) Increasing adoption of data science and machine learning in mainframe environments [2]

(b) Critical need for automated problem resolution to reduce MTTR

(c) Evolution of mainframe modernization strategies [11,12,13]

2. Integration of Modern Tools and Methods

(a) Emergence of open-source data science tools in mainframe ecosystems

(b) Growing adoption of REST APIs for mainframe data access [8]

(c) Need for standardized data preparation methodologies for machine learning applications

(d) Integration of user experience methods for improved stakeholder alignment [3]

3. Data Quality Challenges

(a) 91% of studied mainframe professionals report data quality issues

(b) Persistent challenges in network security problem resolution [6]

(c) Critical need for reliable data in banking operations

(d) Impact of data-driven and communication-driven failures on MTTR [3]

The tutorial directly addresses these challenges by:

- Providing practical methods for improving data quality
- Demonstrating integration of modern tools with mainframe systems
- Teaching effective stakeholder communication techniques
- Presenting real-world examples from banking sector applications

This comprehensive approach supports the mainframe community's evolution toward data-driven operations while maintaining the reliability and security requirements of large-scale enterprise computing [7].

4 Format and Duration

This tutorial consists of a 40-minute lecture that includes a guided demonstration. The lecture provides a tools-based approach to solving data quality challenges in mainframe environments.

- Format: lecture, demonstration, code notebook
- Intended duration: 40 minutes

5 Tutorial Outline

Tutorial Outline (40 minutes):

- Introduction (5 min)
  - Research context and findings
  - Audience background assessment
- Core Concepts (10 min)
  - Data quality in mainframe environments
  - Stakeholder alignment methods
- Hands-on Demonstration (15 min)
  - Data wrangling with VS Code/Jupyter
  - Mainframe API integration
  - Data visualization techniques
- Summary and Implications (5 min)
  - Best practices
  - Implementation strategies
- Q&A Session (5 min)

Note: Timing may be adjusted based on audience needs and technical requirements.

6 Key Learning Objectives

The participants can expect to learn the following:

- **Stakeholder Alignment:** Participants will learn to apply user experience research methods to effectively identify, communicate, and validate business requirements for mainframe data initiatives, reducing communication-driven failures in network security problem resolution.
- **Data Quality Enhancement:** Participants will gain hands-on experience using VS Code, Jupyter Notebooks, and Python to prepare, explore, and transform mainframe network data through REST APIs, addressing the data quality issues reported by 91
- **Practical Implementation:** Participants will learn to apply data wrangling techniques and visualization methods using a combination of mainframe APIs and open-source data science tools.

7 Presenter's Bio

Joshua Powell, Staff Software Engineer at Broadcom's Mainframe Software Division, brings extensive experience in enterprise data systems and artificial intelligence to mainframe modernization. His work spans government agencies, Honeywell AI's Innovation team, and founding the data science venture Viable Industries. At Broadcom, Joshua leads research on improving mainframe data quality and accessibility, in addition to conducting studies with mainframe customers to improve onboarding experiences. His current focus combines user experience methodologies with data science to reduce friction in the adoption of modern mainfra me systems.

8 Tutorial History

The tutorial builds upon an earlier presentation prepared for:

1. 2023 Mainframe Technical Exchange Virtual
2. 2021 Mainframe Technical Exchange Virtual

9 Technical Requirements

Audience members wishing to follow along should prepare their laptops with an up to date version of Visual Studio Code (Microsoft v2024+), Jupyter (Microsoft v2024+), Python v3+, and Python Libraries including Requests, Pandas, and MatPlotLib.

10 Acknowledgments

This work has been entirely supported by the Broadcom Mainframe Software Division.

References

[1] IBM (International Business Machines Corporation). (2023). IBM 2023 Annual Report.

[2] Wishart-Smith, H. (2024, November 13). Mainframes: the backbone of the worldwide economy. Forbes.

[3] Ryseff, J., de Bruhl, B., & Newberry, S. J. (2024). The Root Causes of Failure for Artificial Intelligence Projects and How They Can Succeed: Avoiding the Anti-Patterns of AI.

[4] IBM z/OS operating system. (Accessed: October 28, 2024). https://www.ibm.com/products/zos

[5] Mcgregor, S. E. (2022). Practical Python Data Wrangling and Data Quality. http://oreilly.com

[6] Powell, J.I., Broadcom Mainframe Software Division, Internal Study, April 2024

[7] Alam, A., Bales, R., Dumir, V., Kunze, N., Li, J., Mishra, S., Rivera, E., Wan, M., & Yu, Y. (2024). Turning Data into Insight with Machine Learning for IBM z/OS (First). International Business Machines Corporation.

[8] Broadcom Mainframe Developer Portal. (Accessed: October 28, 2024). https://integration.mainframe.broadcom.com/

[9] Harrell, M. (2024). Mainframe Application Developer Study.

[10] Kanvar, V., Tamilselvam, S., & Raghunath, K. N. (2024, August 8). Enabling communication via APIs for mainframe applications. arXiv.org. https://arxiv.org/abs/2408.04230

[11] Dau, A. T., V., Dao, H. T., Nguyen, A. T., Tran, H. T., Nguyen, P. X., & Bui, N. D. Q. (2024, August 5). XMainframe: a large language model for mainframe modernization. arXiv.org. https://arxiv.org/abs/2408.04660

[12] Raju, J., Modernizing Mainframe Workloads in Banking: Embracing the Power of Hyperscalers, International Journal of Computer Engineering and Technology (IJCET), 15(5), 2024, pp. 366-374.

[13] Raju, J., AI-Driven Transformation of Mainframe Environments: A Comprehensive Framework for Operational Resilience, International Journal of Engineering and Technology Research (IJETR), 9(2), 2024, pp. 420–433.

11 Disclaimer

Certain information in this presentation may outline Broadcom's general product direction. This presentation shall not serve to (i) affect the rights and/or obligations of Broadcom or its licensees under any existing or future license agreement or services agreement relating to any Broadcom software product; or (ii) amend any product documentation or specifications for any Broadcom software product. This presentation is based on current information and resource allocations as of November 6, 2024, and is subject to change or withdrawal by Broadcom at any time without notice. The development, release and timing of any features or functionality described in this presentation remain at Broadcom's sole discretion.

Notwithstanding anything in this presentation to the contrary, upon the general availability of any future Broadcom product release referenced in this presentation, Broadcom may make such release available to new licensees in the form of a regularly scheduled

major product release. Such release may be made available to licensees of the product who are active subscribers to Broadcom maintenance and support, on a when and if-available basis. The information in this presentation is not deemed to be incorporated into any contract.

THIS PRESENTATION IS FOR YOUR INFORMATIONAL PURPOSES ONLY. Broadcom assumes no responsibility for the accuracy or completeness of the information. TO THE EXTENT PERMITTED BY APPLICABLE LAW, BROADCOM PROVIDES THIS DOCUMENT "AS IS" WITHOUT WARRANTY OF ANY KIND, INCLUDING, WITHOUT LIMITATION, ANY IMPLIED WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE, OR NONINFRINGEMENT. In no event will Broadcom be liable for any loss or damage, direct or indirect, in connection with this presentation, including, without limitation, lost profits, lost investment, business interruption, goodwill, or lost data, even if Broadcom is expressly advised in advance of the possibility of such damages.