Ohio Department of Higher Education | Ohio Education Research Center

A Case Study on Ensuring Data Quality Across a Multi-Source Pipeline: Evaluating Tableau Dashboard Built from Oracle Data and Legacy PDF Reports for Educator Preparation Program

Floria Liu

I. Introduction

- **What is Data Quality?**

Data quality refers to how well data meets the requirements of its intended use. It is commonly assessed through key dimensions such as accuracy, consistency, completeness, and timeliness (Wang, Liu, Li, & Lin, 2024). These dimensions help determine whether data can be trusted for decision-making and reporting.

- **Why Should We Care?**

Inaccurate or inconsistent data can result in misleading dashboards, misinformed decisions, and lost confidence among stakeholders. DCI Wealth (2023) estimates that poor data quality costs organizations an average of $12.9 million per year, and that unresolved errors multiply in cost the further they move downstream in the data pipeline.

- **What Does Good Data Look Like?**

Good data is both accurate and consistent. Wang et al. (2024) emphasize that high-quality data should be structurally and semantically reliable, especially in systems involving multiple teams and formats.

II. Conceptual Foundations

1. Key Concepts

Concept	Definition
Accuracy	Data correctly reflects the real-world value it represents
Consistency	Data values are presented in the same format across systems or time

*Definitions based on Wang et al. (2024).

2. Tableau Data Field

Concept	Meaning
Dimension	Qualitative labels used to group or categorize data *e.g., program name, license year*
Measure	Quantitative values used in calculations *e.g., number of completers, pass rate*

III. Case Study

1. Background

Ohio Revised Code § 3333.048 requires the Chancellor of Higher Education and the Superintendent of Public Instruction to establish and publish performance metrics for educator preparation programs. In response, the Ohio Department of Higher Education (ODHE) collaborates with the internal & external agencies, and higher education institutions to collect and report data on key outcomes. This project explored the data preparation pipeline for creating and publishing the Educator Preparation Performance Dashboard on the ODHE website. The goal was to identify areas where greater accuracy and consistency could support ODHE in delivering these reports more reliably.

2. Methodology

- Queried Oracle database using SQL to extract relevant data
- Used Excel to review the original submitted data files and validate them against what was stored in Oracle
- Compared Oracle results to the legacy PDF report and Tableau dashboard
- Assessed data quality by checking for:
  i. Accuracy – differences in values across sources
  ii. Consistency – mismatched field names or labels

3. Ohio Educator Preparation Program data pipeline chart

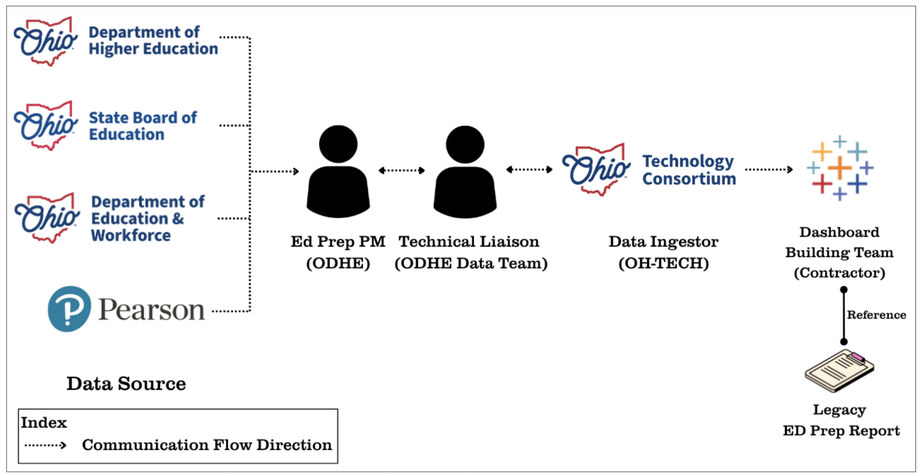

Figure 1. Ohio Educator Preparation Program data pipeline chart

4. Process Flow Explanation

Data comes from third-party vendors, state agencies, and ODHE's internal sources. The Technical Liaison helps the Ed Prep Project Manager prepare these files, which are then uploaded to a secure portal and placed into the Oracle server by OH-TECH. The Dashboard Building Team queries the Oracle server to build visualizations, using a legacy PDF report as a layout reference. They do not modify or upload any data. The legacy PDF report is publicly available; the Oracle database is not publicly accessible. The Tableau dashboard is expected to be made publicly available in the future.

5. Example of Accuracy & Consistency Issues

i.Legacy Ed Prep Report (Teacher Resident Program)

Initial Licensure Effective Year	Residency Year 1		
	Entering	Persisting	
2019	79	77	97.5%

ii.Database

INST_NAME	LICENSE_YEAR	YEAR_1_ENTERING	YEAR_1_COMPLETING
Ohio Department of Higher Education	2019	79	77

iii.Tableau Dashboard (Teacher Residency)

	Residency Year 1		
Initial Licensure Effective Year	Entering	Persisting	Avg. Pct
2019	158	154	97.5%

Explanation:

- Inaccuracy: The number of candidates "Entering" and "Persisting" for 2019 in Residency Year 1 differs between sources — the legacy report and database both show 79/77, while the Tableau dashboard shows 158/154, nearly double the correct counts.
- Inconsistency: The same metric is labeled as "Persisting" in the legacy report and Tableau, but appears as "Completing" in the database. This inconsistency in terminology makes tracking and comparison difficult across systems.

IV. Conclusion

Ensuring data quality is not just a technical task, it requires aligned workflows, role clarity, and shared accountability across teams. This case study demonstrates that without shared understanding and data fluency, even well-intentioned systems can produce misaligned results. Strengthening collaboration and embedding quality checks early can help organizations transform reactive troubleshooting into proactive data stewardship. The table below shows the specific challenges and recommendations for this case:

Challenge	Recommendation / Best Practice
Limited accessible documentation on legacy extraction and Oracle uploads	Require schema documentation and upload logs for every data transfer
Ununified labeling across data sources	Standardize program naming conventions across systems
Discrepancies across multiple fields between Tableau and legacy PDF reports	Implement validation checkpoints between Oracle data and Tableau visual outputs
Distributed ownership with limited cross-team visibility	Define clear role boundaries and assign data quality oversight responsibilities
Team members specialize in different fields, not all in data management	Offer basic data training or include a dedicated data steward in the workflow

V. Discussion

This case study reveals how unclear documentation, ununified labels, and siloed responsibilities can lead to data inconsistencies in public-facing dashboards. These issues reduce trust in the system and make troubleshooting difficult across teams.

$ → $$ → $$$

Prevention → Resolution → Correction

As mentioned in the DCI Wealth (2023), correcting data errors after publication can cost up to 100 times more than preventing them at the source. Proactive practices like standardized naming and upload tracking are essential for reducing long-term risks and improving reporting quality.

VI. Acknowledgements

I would like to thank Kori Khan for her guidance and support as my supervisor throughout this internship. Special thanks to Josh Hawley and Ceanna Burnheimer for their consistent weekly support and encouragement. I also appreciate Kira Steigerwald Johns for helping me practice and refine my presentation, and Hongji Chen for supporting the other project I was involved in during this time. I'd also like to thank my peers, Zackary Howes, Killian Hoyt, and Yaqi Zhang, for their collaboration and shared learning throughout the experience.

VII. Reference

DCI Wealth. (2023, January). *The hidden cost of bad data*. wealth-dci.com. https://www.wealth-dci.com/wp-content/uploads/2023/01/dci-whitepaper-the_hidden_cost_of_bad_data.pdf

Wang, J., Liu, Y., Li, P., Lin, Z., Sindakis, S., & Aggarwal, S. (2024). Overview of Data Quality: Examining the Dimensions, Antecedents, and Impacts of Data Quality. *Journal of the Knowledge Economy*, *15*(1), 1159–1178. https://doi.org/10.1007/s13132-022-01096-6

Ohio Department of Higher Education. *2023 Ohio Educator Preparation Provider Performance Report, State Report.*

THE OHIO STATE UNIVERSITY