Journal Publication of International Research for Engineering and Management (JOIREM)
Volume: 03 Issue: 11 | Nov-2025
ISSN (O) 3107-6696

MINI LAKEHOUSE ON DUCKDB + LOOKER STUDIO: A DATA ENGINEERING PIPELINE

Ms. Saniya Shafi Ahmed Shaikh, Prof. A. S. Sardar, Dr. S. G. Sahani, Dr. S. P. Abhang, Prof. P. S. Umate, Dr. S. V. Khidse

[1]M. Tech Student of Department of Computer Science & Engineering, CSMSS Chh. Shahu College of Engineering,
[2]Prof. at Department of Computer Science & Engineering, CSMSS Chh. Shahu College of Engineering, Chhatrapati Sambhajinagar, Maharashtra, India.

Abstract - Small organizations often rely on messy, inconsistent spreadsheets that limit analytics quality. This paper presents a Mini Lakehouse architecture built using DuckDB, Parquet, and Python, with Looker Studio dashboards. The workflow ingests Excel data, performs structured cleaning, builds a star schema, enforces data quality checks, and stores curated data in Parquet/DuckDB. This fully local, low-cost pipeline provides reproducible, auditable analytics without cloud infrastructure. The result is a governed, BI-ready system suitable for small teams and academic environments.

Key Words: Data Pipeline, Lakehouse, DuckDB, Parquet, Looker Studio, Star Schema, Data Quality, Dashboards

1. INTRODUCTION

Small organizations commonly depend on Excel-based reporting, creating inconsistencies and limiting analysis. This project implements a Mini Lakehouse architecture using Parquet storage and DuckDB for fast SQL analytics. The system cleans raw Excel data, standardizes formats, and builds BI-friendly dimensional models.

Literature Review

Lakehouse systems unify data warehouses and lakes. DuckDB provides in-process OLAP performance, while Parquet offers compressed columnar storage. Dimensional modeling simplifies BI workflows and supports governed analytics.

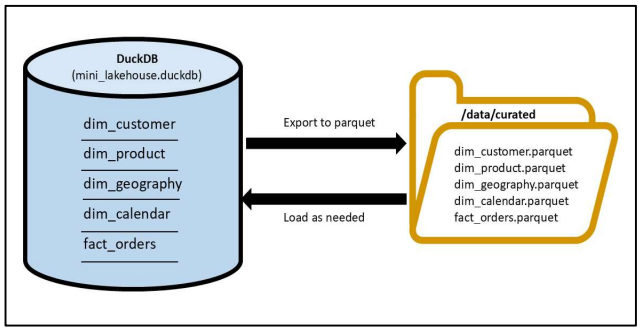
METHODOLOGY / SYSTEM ARCHITECTURE

The architecture includes Raw, Staging, and Curated zones. Cleaning includes type coercion, null handling, and category normalization. The Curated zone models a Fact Orders table linked to Customer, Product, Geography, and Calendar dimensions.

4. IMPLEMENTATION

Implemented using Python notebooks and DuckDB. Data is ingested from Excel, cleaned, transformed, and stored as Parquet. Incremental loads use hash-based surrogate keys. Benchmarks show significant performance gains.

Figure-1: Physical Storage Mapping



5. RESULTS AND DISCUSSION

The curated schema supports fast queries and Looker Studio dashboards. Data quality checks validate referential integrity and completeness. Dashboards deliver insights across segments, geographies, and time.

6. CONCLUSION

The Mini Lakehouse demonstrates a portable, low-cost, reproducible analytics pipeline. It supports governance, incremental updates, and BI dashboards without cloud dependencies.

REFERENCES

[1] Hai, R. et al., Data Lakes: A Survey of Functions and Systems.

[2] Azzabi, S. et al., Data Lakes: A Survey of Concepts and Architectures.

[3] Armbrust, M. et al., Delta Lake: High-Performance ACID Table Storage.

[4] Kohn, A. et al., DuckDB-Wasm: Fast Analytical Processing.

[5] Liu, C. et al., Performance Analysis of Columnar Formats.

[6] Mezzoudj, S., et al. (2025). "Data Lakes versus Data Warehouses: Choosing the Right Architecture." Journal of Engineering and Applied Science (SpringerOpen). SpringerOpen 9 CSMSS,Chh Shahu College of Engineering

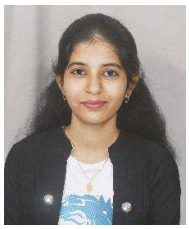
[7] Bernardo, B. M. V., et al. (2024). "Data Governance & Quality Management—Innovation and Research Trends: A Comprehensive Review." (Elsevier journal). ScienceDirect

[8] Arundel, S. T., et al. (2023). "A Guide to Creating an Effective Big Data Management Framework." Journal of Big Data (SpringerOpen). SpringerOpen

[9] Shojaee Rad, Z., et al. (2024). "Data Pipeline Approaches in Serverless Computing." Journal of Big Data (context: lake-backed pipelines). SpringerOpen

[10] Ivanov, T., et al. (2020). "The Impact of Columnar File Formats on SQL-on-Hadoop Performance (incl. Parquet)." Concurrency and Computation: Practice and Experience (Wiley). Wiley Online Library

BIOGRAPHIES



Ms. Saniya Shafi Ahmed Shaikh,
M. Tech Student of Department of Computer Science & Engineering, CSMSS Chh. Shahu College of Engineering,Chhatrapati Sambhajinagar, Maharashtra, India.

Prof. A. S. Sardar
Prof. at Department of Computer Science & Engineering, CSMSS Chh. Shahu College of Engineering, Chhatrapati Sambhajinagar, Maharashtra, India.

Dr. S. P. Abhang
Prof. at Department of Computer Science & Engineering, CSMSS Chh. Shahu College of Engineering, Chhatrapati Sambhajinagar, Maharashtra, India.

Prof. P. S. Umate
Prof. at Department of Computer Science & Engineering, CSMSS Chh. Shahu College of Engineering, Chhatrapati Sambhajinagar, Maharashtra, India.

Dr. S. V. Khidse
Prof. at Department of Computer Science & Engineering, CSMSS Chh. Shahu College of Engineering, Chhatrapati Sambhajinagar, Maharashtra, India.