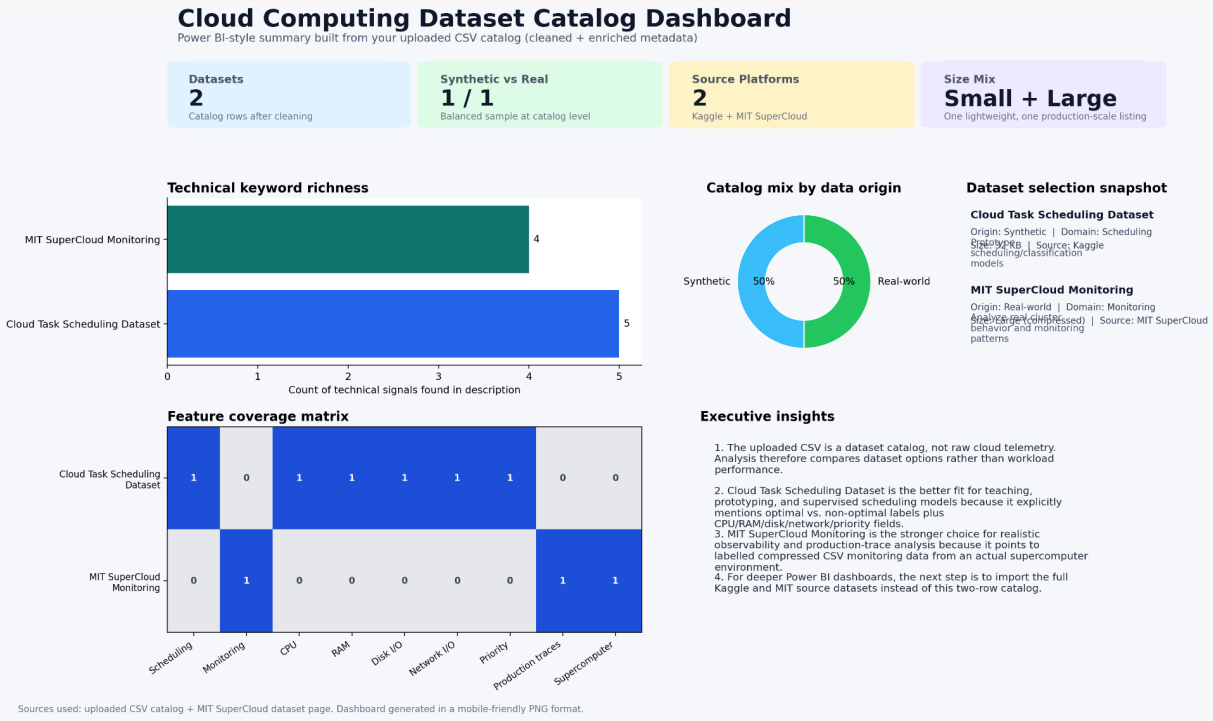


Cloud Computing Dataset Catalog



Overview

Power-BI-style catalog comparing two cloud computing data sources — one synthetic (Kaggle) and one real-world (MIT SuperCloud) — to guide dataset selection for scheduling and observability work.

Key Findings

- Balanced sample: 1 synthetic / 1 real-world source.
- Cloud Task Scheduling Dataset: best fit for supervised scheduling models (CPU/RAM/disk/network/priority fields).
- MIT SuperCloud Monitoring: strongest choice for realistic production-trace analysis.
- Feature coverage matrix surfaces which dataset answers which question.

Business Value

Saves teams weeks by choosing the right dataset before writing a single line of code.