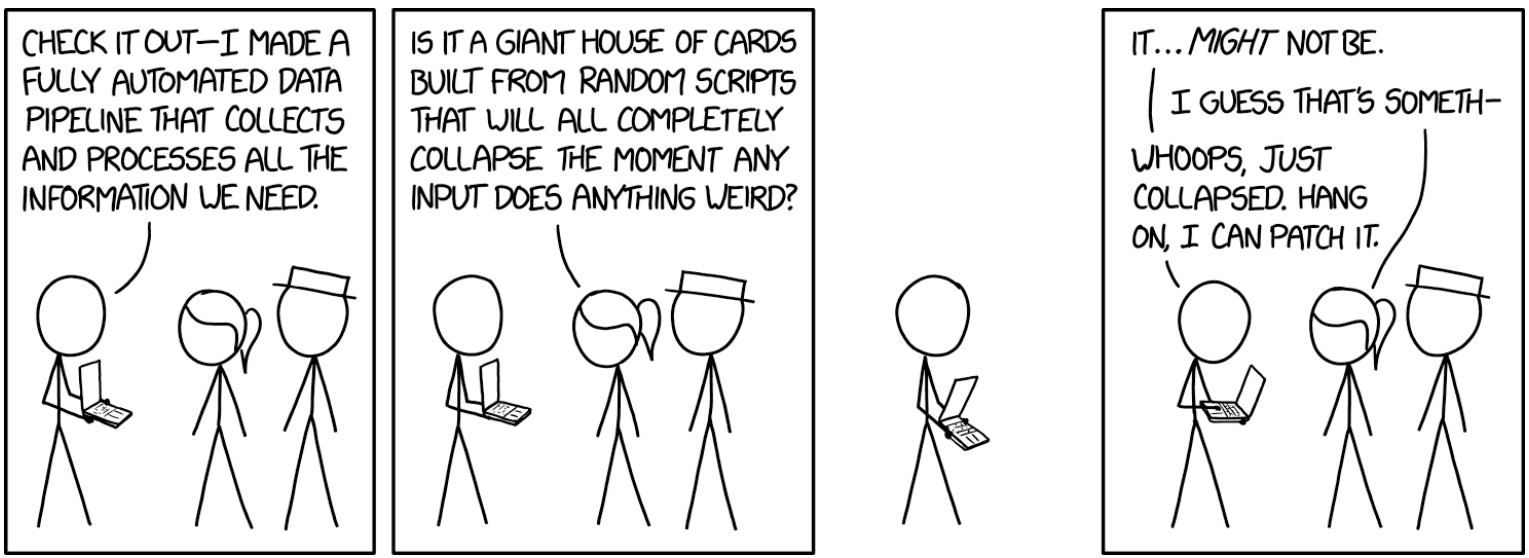
Data & Analytics Case Study

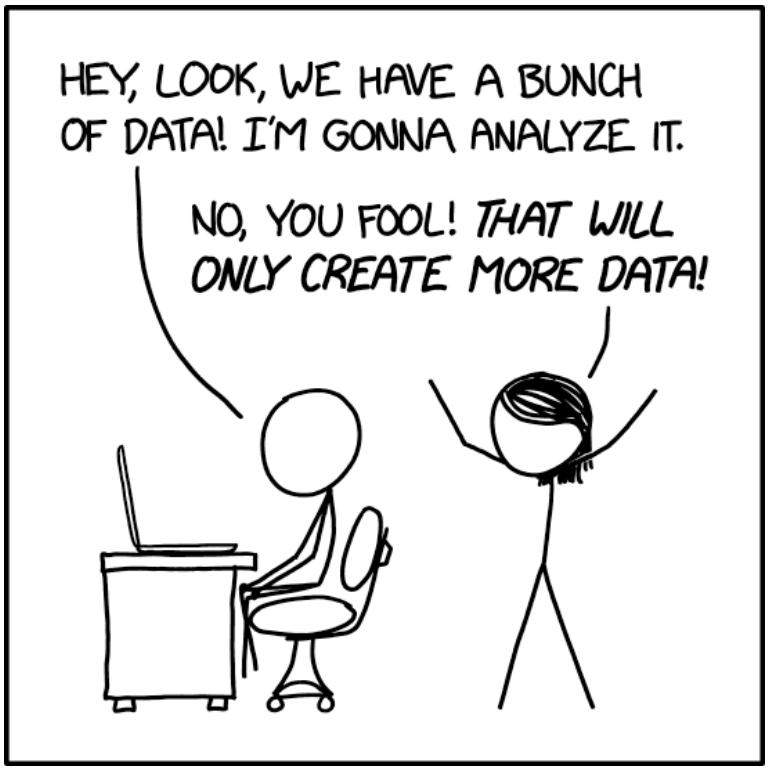
Max van Rooijen (max@pl9.co)

March 11, 2025



Why Analytics in Comic Design?

- Understanding audience preferences and engagement
- Optimizing storytelling and visual elementsd, see what works and what doesn't
- Enhancing content based on costs, views and reviews



Technical Solution: Batch Processing Pipeline

The How

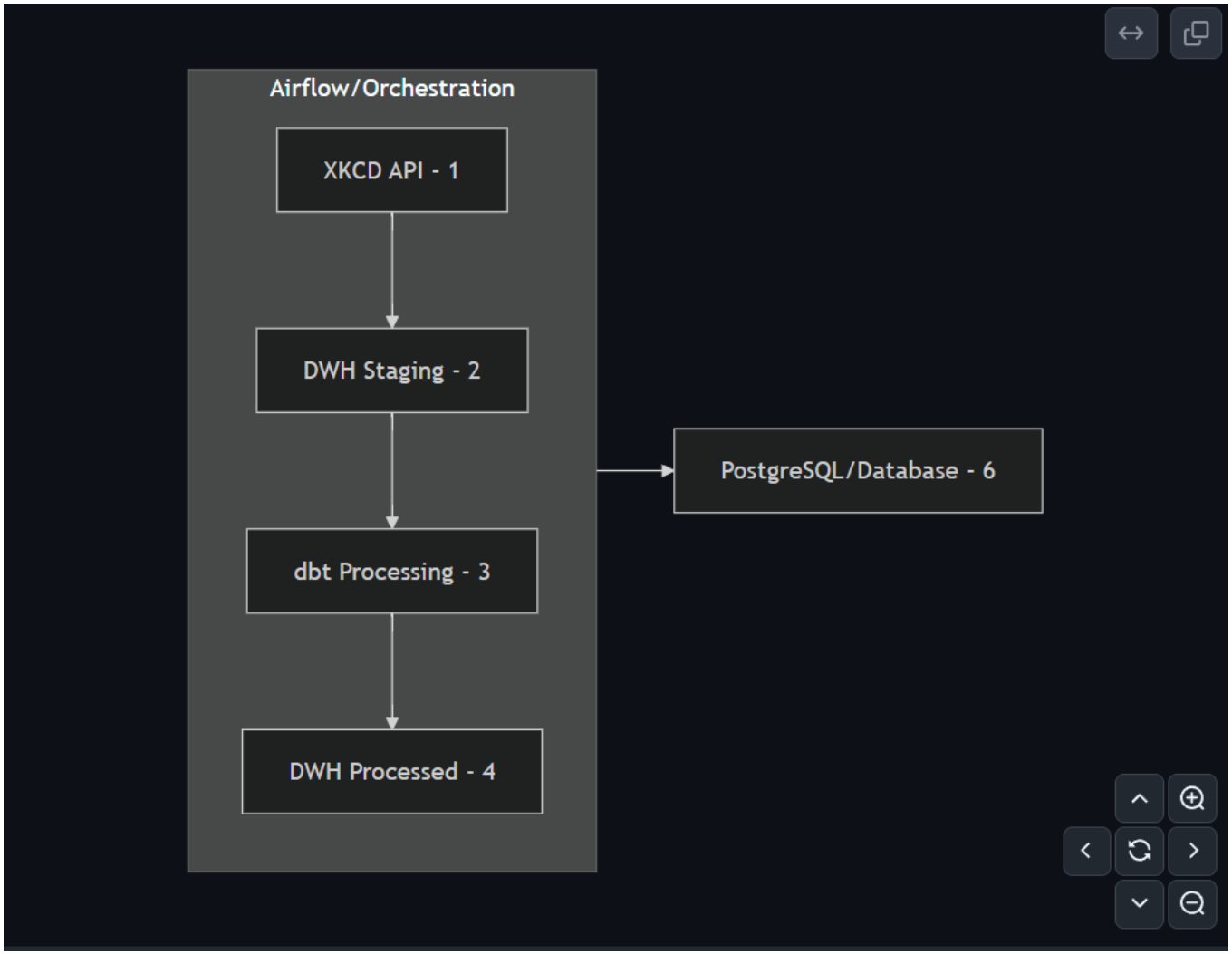
The overall data pipeline executes these tasks:

- Fetching data from the XKCD API
- Using Apache Airflow to orchestrate workflows
- Staging data in a PostgreSQL Data Warehouse
- Transforming raw data using *dbt*
- Running data quality checks to ensure integrity

Technologies used:

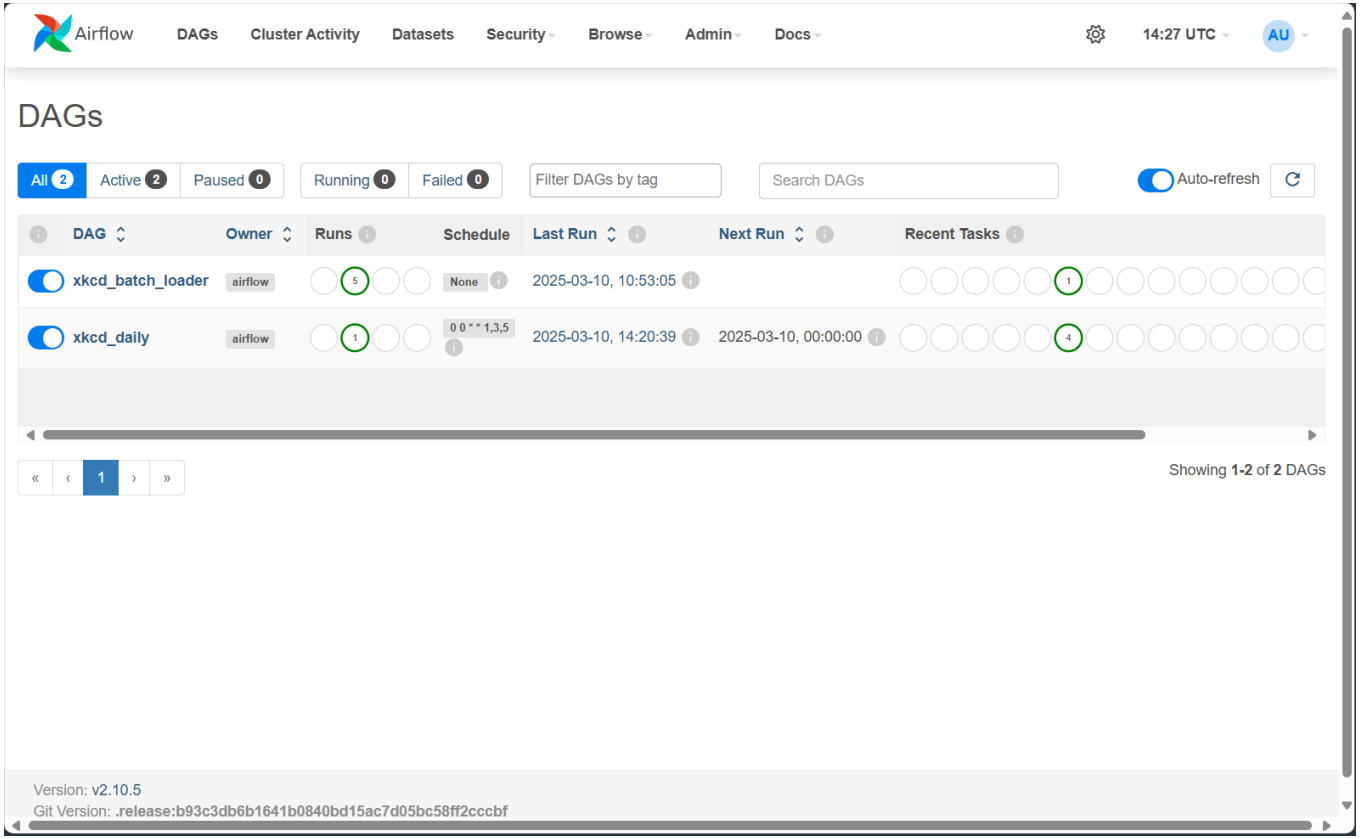
- **Apache Airflow**: To orchestrate workflows.
- **PostgreSQL**: For staging data in a Data Warehouse.
- **dbt (data build tool)**: For transforming raw data.

Solution Overview



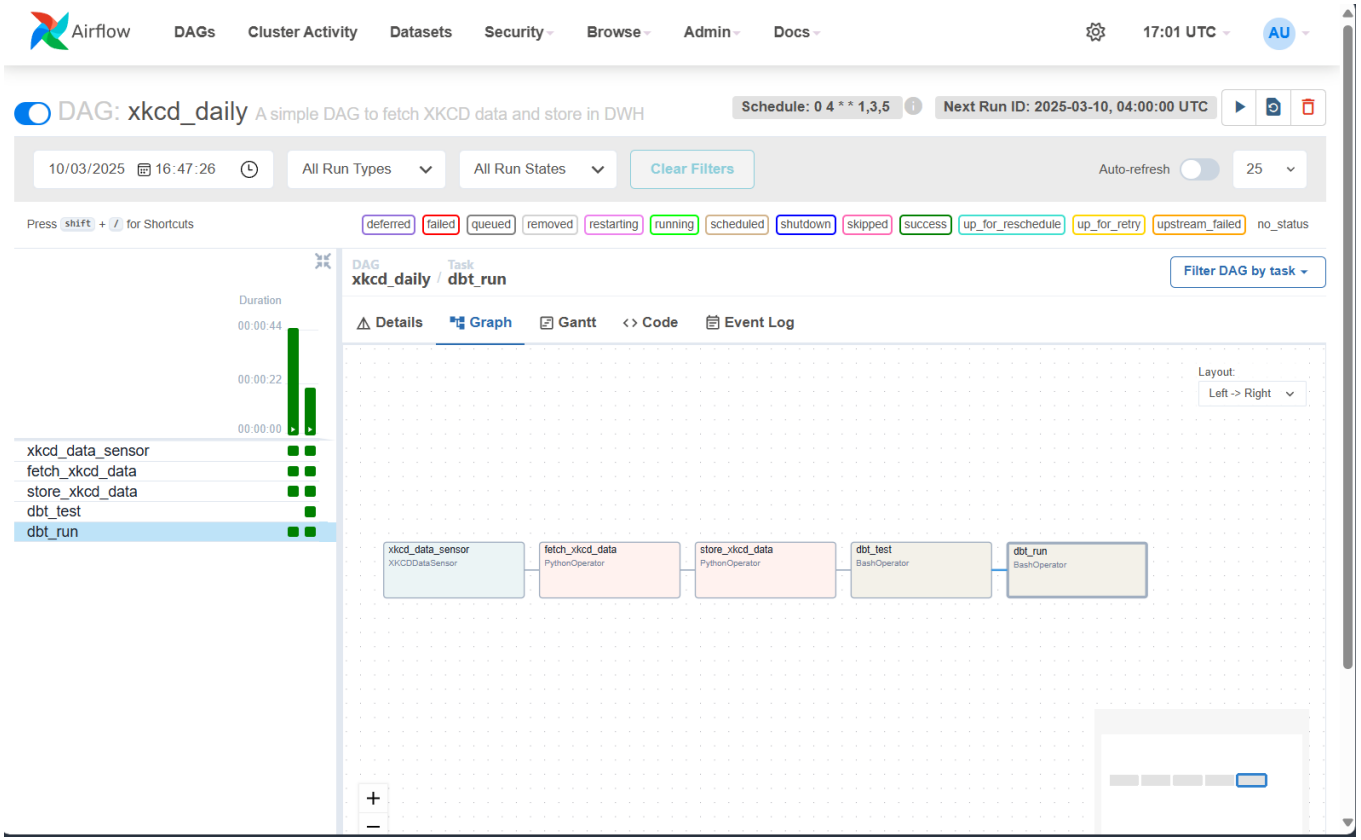
High-level overview of the architecture.

Pipeline Architecture



Overview of the DAGs in Airflow

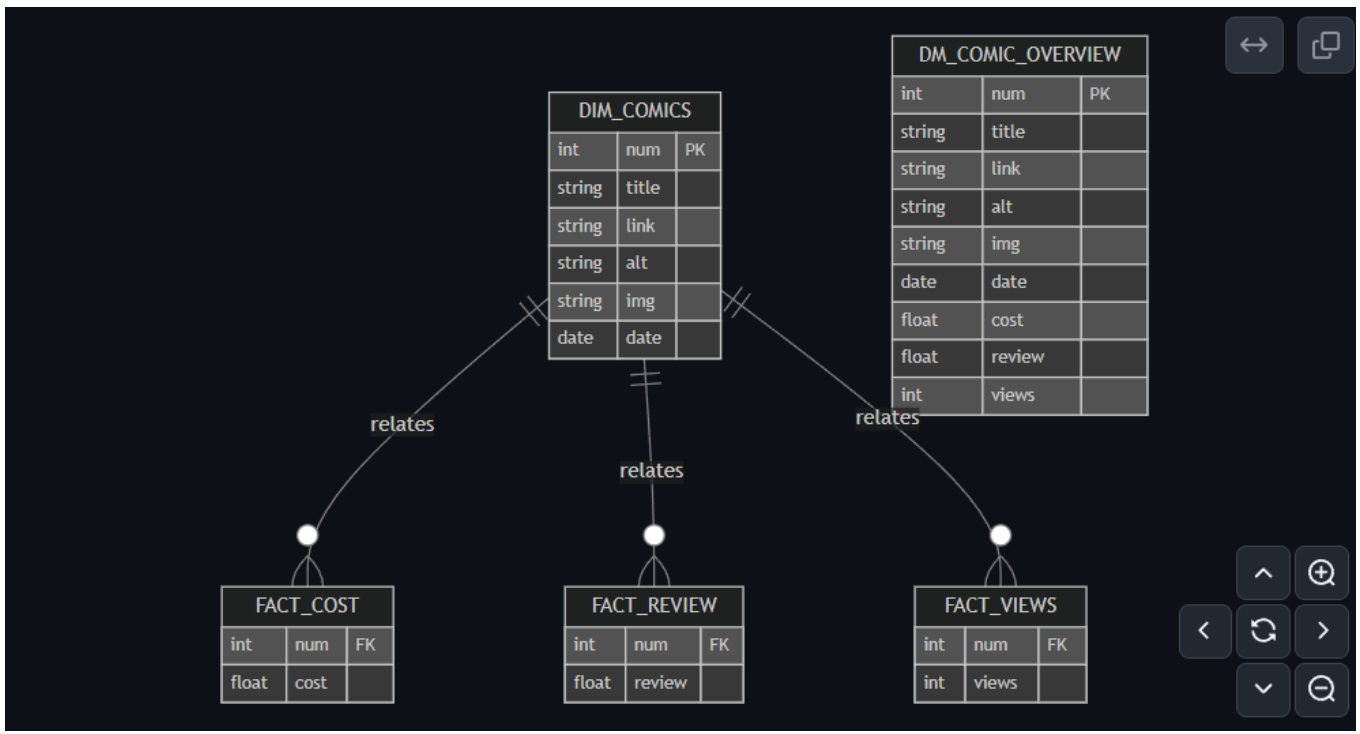
Pipeline Architecture



Daily run DAG

Data Model and Insights

- Dimensional model for structured data storage
- Fact tables for costs, reviews, and views
- Aggregated insights for trend analysis

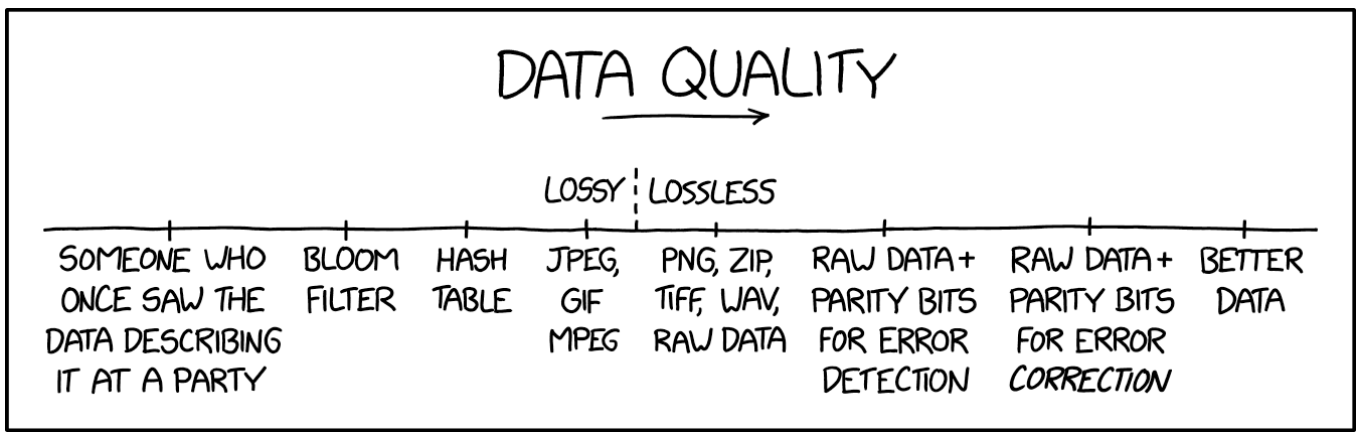


Future Technical Improvements

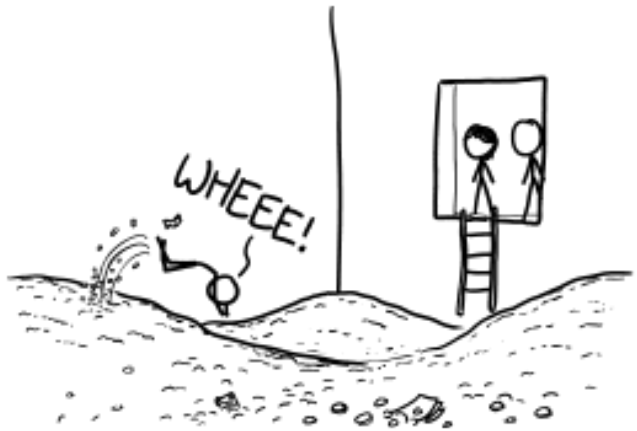
- **Machine Learning Integration**: Use ML models for predictive insights, such as forecasting trends in comic engagement.
- **Enhanced Data Quality Checks**: Implement more comprehensive tests for duplicates, referential integrity, and outlier detection.
- **Unit Tests for DAGs**: Add unit tests to Airflow DAGs to ensure robust workflow execution.
- **More Elaborate Data Model**: Expand the data warehouse schema to include more detailed metadata and user behavior analytics.
- **Scalability**: Optimize the data pipeline to handle larger volumes of data and improve performance.
- **Data Lineage**: Implement the dbt docs routine to automatically generate an interactive document in HTML and host it on a web server.

Research Questions

- Which comics have the highest engagement rates and why?
- How do different types of content (e.g., tech, society) perform in terms of views and shares?
- What is the cost per engagement for the comics?
- Are there any trends in reader preferences over time?
- What are the common themes in reader feedback and reviews?
- How can we optimize our content strategy to increase reader retention and engagement?
- How does reader engagement correlate with revenue from ads and t-shirt sales?



Future Business Cases



- **Optimizing Ad Placement**: Use insights from user engagement to strategically place advertisements in comics.
- **Subscription & Monetization Strategies**: Identify premium content opportunities based on engagement trends.
- **Predicting Viral Content**: Utilize data to forecast which comics are likely to go viral and maximize their exposure.
- **Improving Reader Retention**: Track engagement metrics to refine content strategy and keep readers coming back.