ASSIGNMENT 15: Orchestration & System Design

For my project, a logical breakdown of tasks would be:

- **ingest_data**: Fetches the latest daily market data from the Yahoo Finance source for a specified date range.
- **validate_data**: Checks the raw data for completeness (no missing dates), expected columns, and excessive nulls.
- **engineer_features**: Calculates the necessary metrics from the clean data (e.g., Daily Return %, Volume Change %, Volatility %).
- **run_analysis**: Performs the core regression and classification experiments to get the R-squared and accuracy metrics.
- **generate_report**: Takes the statistical results and creates a summary artifact (e.g., a markdown file with key metrics and plots).
- **check_for_drift**: Compares the new R-squared value against a threshold (e.g., 0.05) to see if the "no relationship" conclusion still holds. If the threshold is breached, it flags an alert.

DAG for Volume-Price Analysis

This diagram shows the simple, linear sequence of tasks in your analysis pipeline. Each step must be completed successfully before the next one begins.

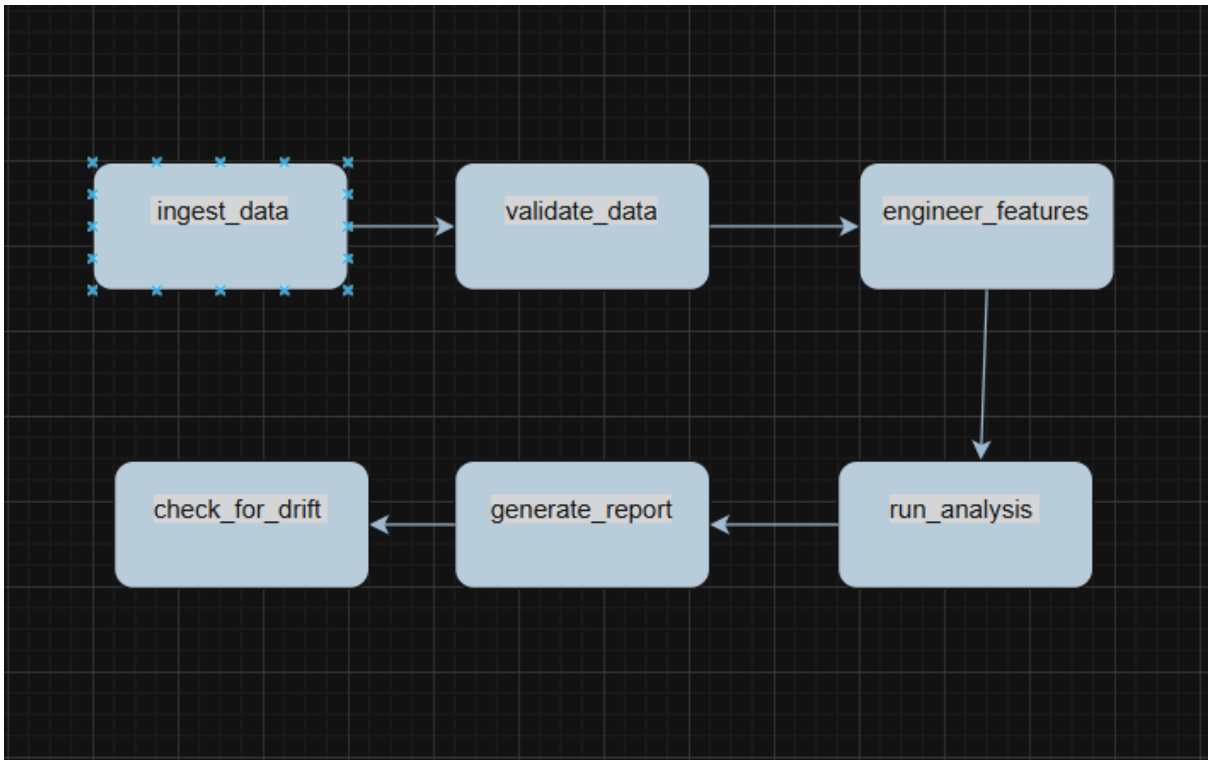

Table that details the inputs, outputs, and reliability strategy for each task.

Task Name	Inputs	Outputs	Idempotent?	Logging & Checkpoints
ingest_data	Date range	data/raw/djia_raw.csv	Yes (overwrites file)	Log rows fetched. Output CSV is the checkpoint.
validate_data	data/raw/djia_raw.csv	data/processed/djia_validated.csv	Yes (overwrites file)	Log validation success/failure. Output is the checkpoint.
engineer_features	data/processed/djia_validated.csv	data/final/djia_features.csv	Yes (overwrites file)	Log features created. Output is the checkpoint.
run_analysis	data/final/djia_features.csv	results/latest_metrics.json	Yes (overwrites file)	Log R-squared and accuracy. JSON output is the checkpoint.
generate_report	results/latest_metrics.json	reports/weekly_report.md	Yes (overwrites file)	Log report generation success.
check_for_drift	results/latest_metrics.json	logs/pipeline.log (entry)	No (appends to log file)	Log the drift check result (e.g., "R-squared below threshold").