The Market is a Conversation. We Built the Platform to Listen. So You Can Profit.

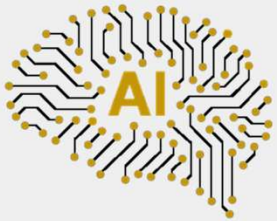
Market movements are increasingly driven by public sentiment, not just traditional financial metrics. To gain an edge, we must leverage on data and find the Alpha.

Team 4:
Pang Soh Har
Pong Chi Leong
Tan Chia Lin (Eunice)
Tjoa Reinhard
Thandar Myo Myint
Lim Lik Loong

Our Mission: To Find Profitable Pattern from Stock Market

We built An end-to-end data platform for transforming raw market signals into actionable insights.

Automated Daily Extraction

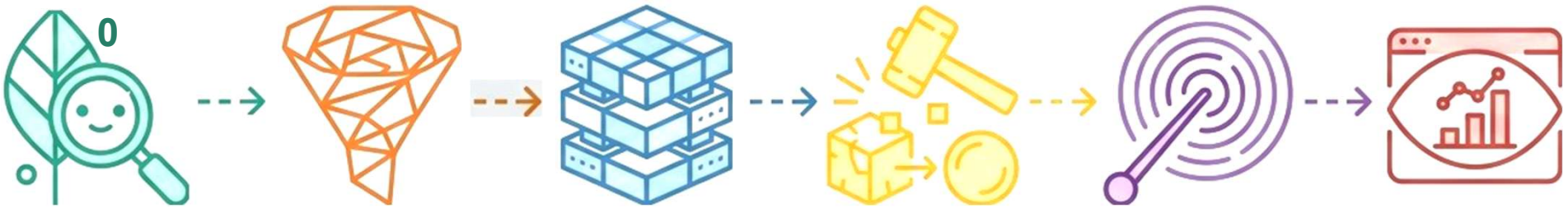


AI-Driven Market Intelligence

Cloud Infrastructure

The Engine Room: A Modern, Modular Data Stack

This is a robust, end-to-end ELT pipeline built on a modern, scalable data stack.



Extraction (Python): Custom scripts act as our data scouts, gathering stock prices, news headlines, and market indices.

Ingestion (Meltano): The universal adapter, pulling raw data from any source into our central repository.

Storage (Google BigQuery): The scalable and reliable core of our Lakehouse, holding all the raw and refined information.

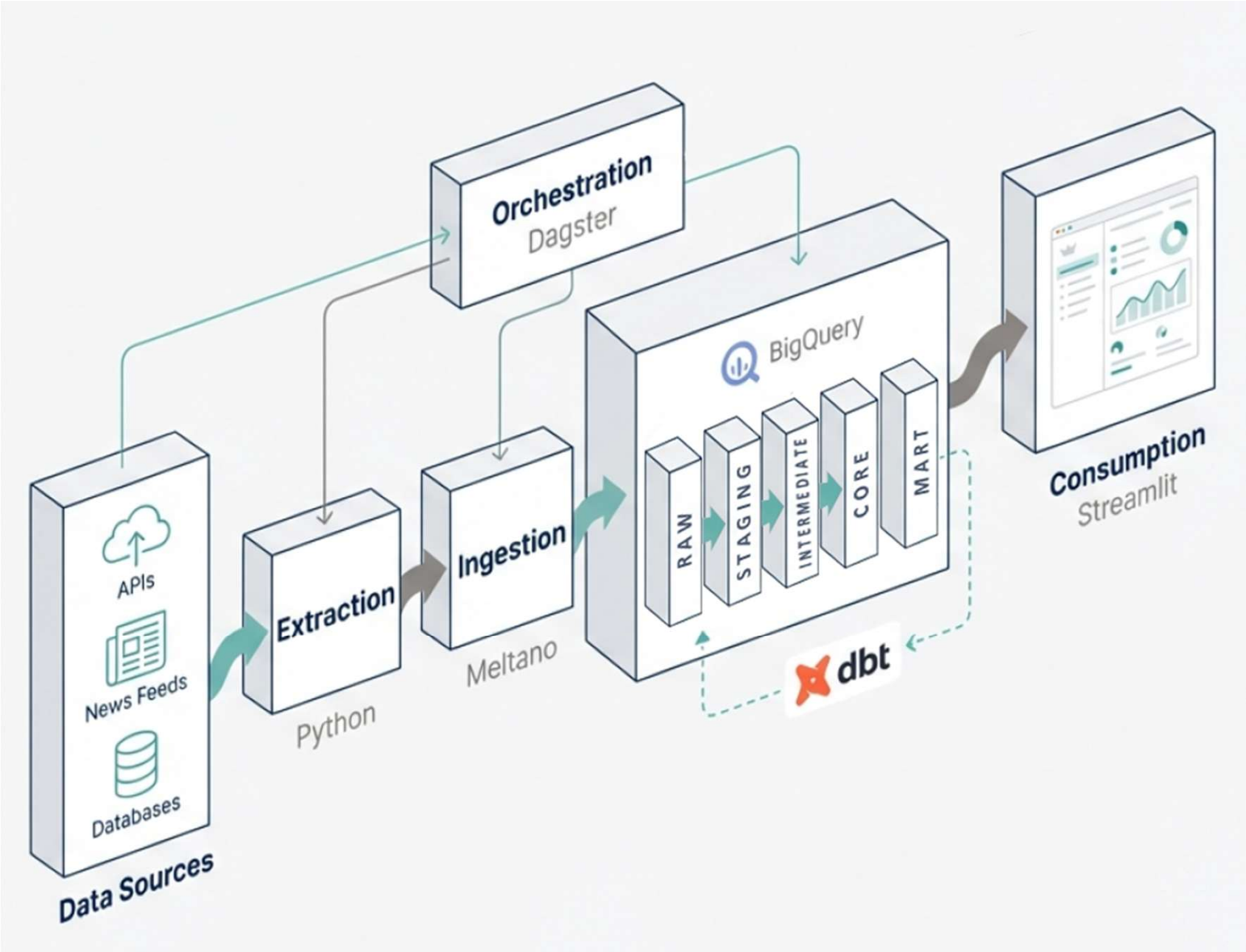
Transformation (dbt): The alchemist, turning raw data into clean, structured, and analytics-ready insights.

Orchestration (Dagster): The conductor, ensuring every part of the pipeline runs flawlessly and on schedule, every day.

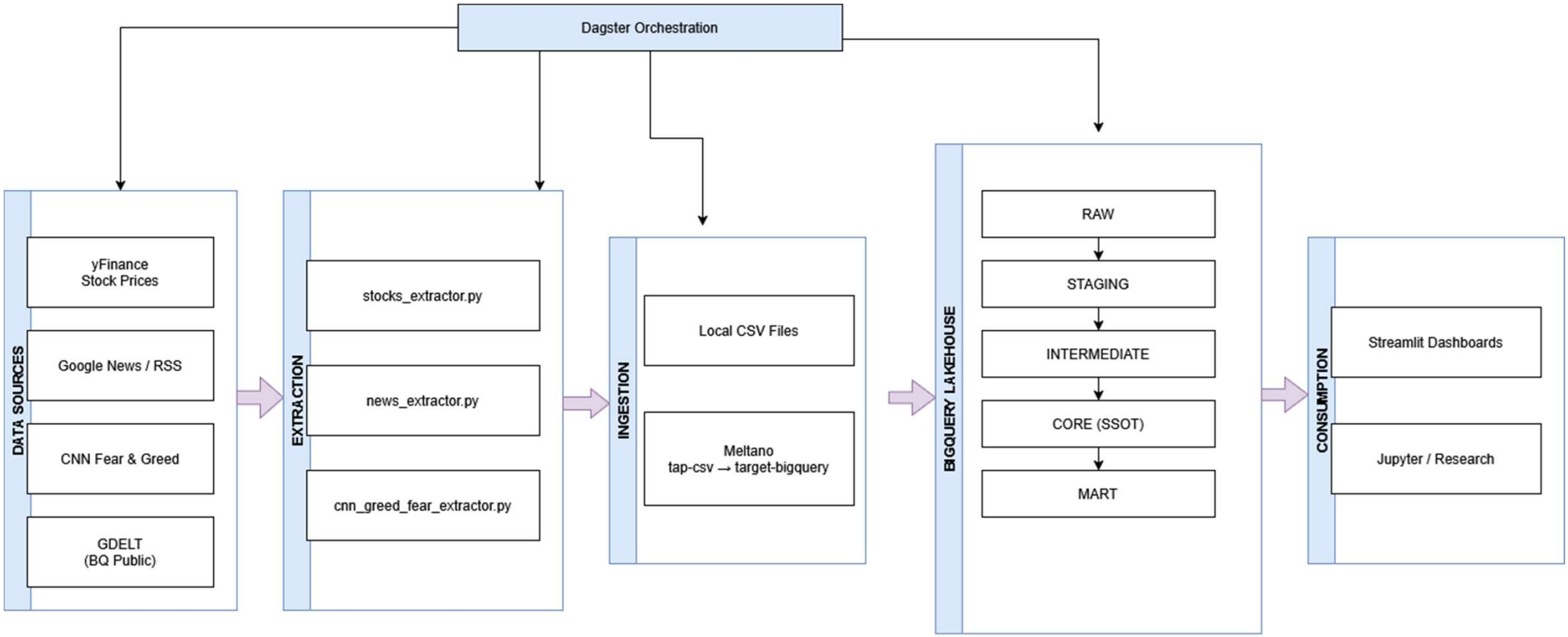
Visualization (Streamlit): The command center, providing an interactive window into the Market Brain's thoughts.

The Blueprint:

From Disparate Data to Integrated Intelligence



Data Flow: Managing Files Across the Pipeline

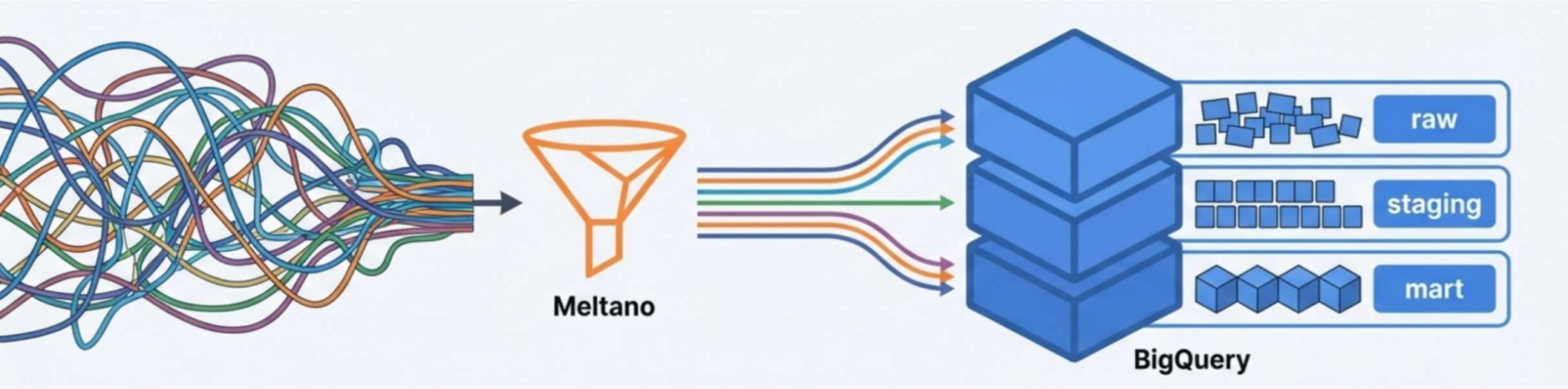


Data Source: Market's Data, News,and Investor Sentiment

The pipeline integrates four primary types of market and sentiment data, captured by dedicated Python extractors.

- **Stock Prices:** Daily OHLCV data for the "Magnificent Seven" stocks, key indices (e.g., S&P 500), and the VIX from 'yfinance'.
- **News Sentiment:** Headlines from Google News, with sentiment analysis applied using a FinBERT model.
- **Market Fear Index:** The daily CNN Fear & Greed Index to gauge broad market sentiment.
- **Global Events:** News event data from the BigQuery public GDELT GKG dataset, used to map events to specific tickers.

The Ingestion Engine: From Chaotic to a Lakehouse



Meltano is used to manage the EL (Extract-Load) process. It uses 'tap-csv' to read the files produced by the Python extractors and 'target-bigquery' to load them into the 'mag7_intel_raw' dataset. This approach decouples extraction logic from loading mechanics.

The dbt Refinery: Structuring Data for Analysis

We employ a multi-layered modelling approach using dbt. Each layer serves a distinct purpose, transforming the data step-by-step to increase its quality, structure, and analytical value. This ensures a clear data lineage and maintainable transformation logic.

We Start from Raw

Raw data is messy, duplicated, and inconsistent. We use dbt to enforce quality and structure

Model Layer: Enriching with Business Logic

Staging Layer

The staging layer takes the raw, source aligned data and applies foundational transformations. This includes type casting, deduplication, and key normalisation. It serves as the single source of truth for cleaned base entities.

Intermediate Layer

we combine and enrich the staging models to create more complex data structures. This is where we add calculated metrics and join different data sources together.

Core Layer

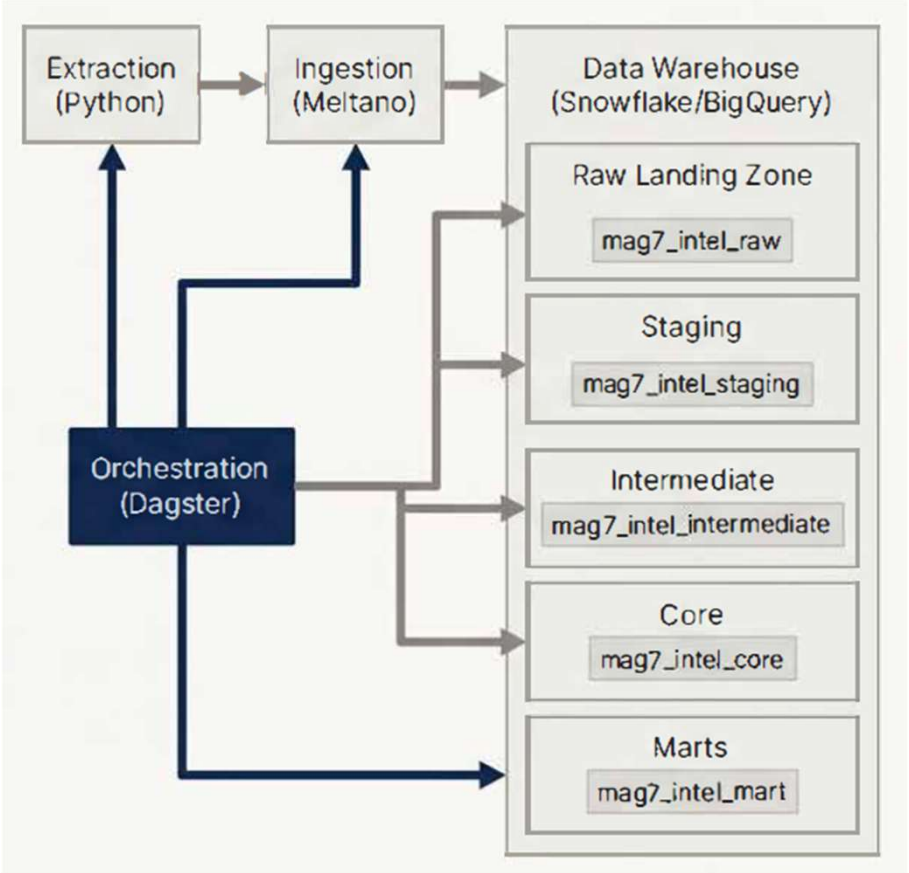
Builds canonical, production-grade tables like fact_prices and fact_price_ features (returns, volatility, momentum). These are robust, well-tested datasets.

Mart Layer

Creates wide, de-normalised tables specifically for consumption. These models directly power dashboards and common analytical queries, like macro_risk_dashboard and ticker _overview.

Orchestrating: Designing and Automating the End-to-End Pipeline

Dagster automates the end-to-end workflow. It defines each step-from Python extraction scripts to Meltano jobs to dbt model runs-as a software-defined asset. This allows Dagster to manage dependencies and trigger runs based on data updates, ensuring the entire pipeline is reliable and observable.

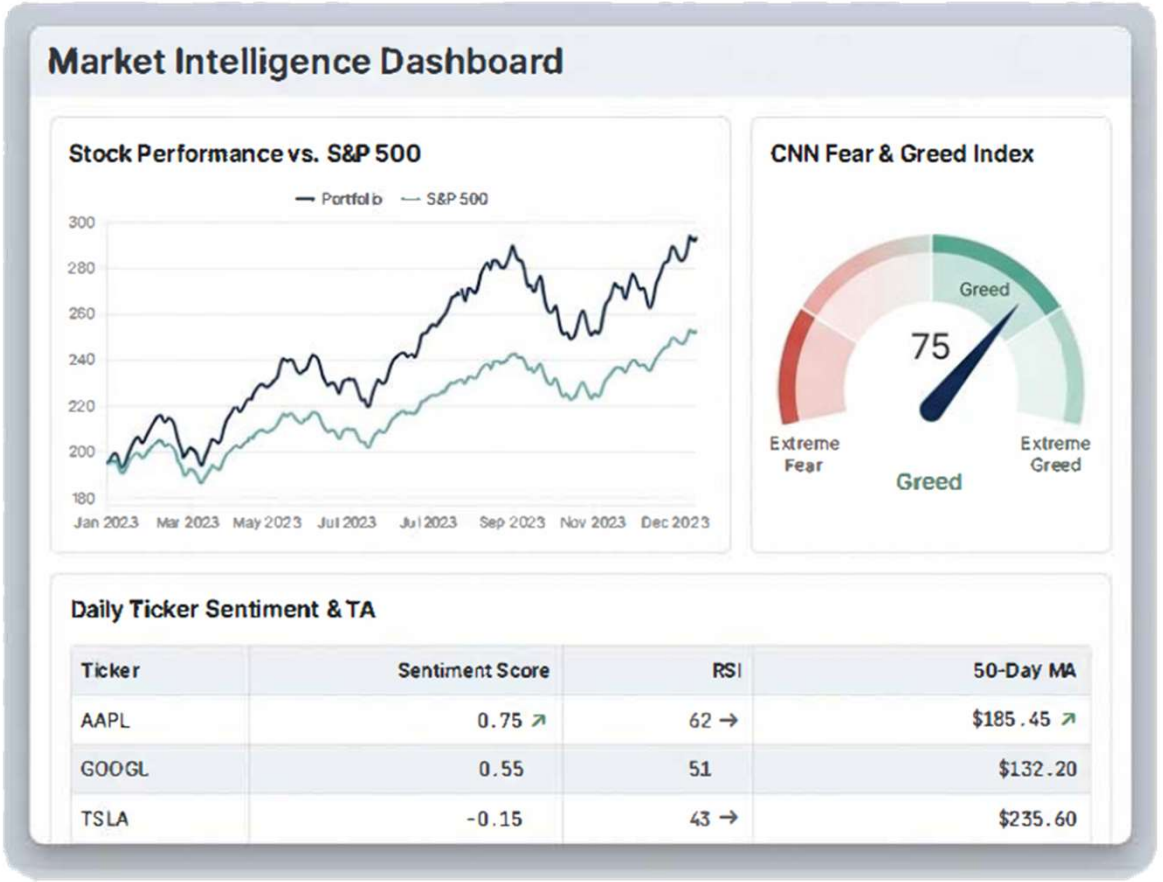


Key Orchestrated Assets

1. Python extractors (run_news_extractor, extract_to_csv)
2. Meltano EL job (mel tano run load_csvs)
3. dbt transformation jobs (dbt run, dbt test)

The Result: Quant Research Interface

The data marts directly power an interactive Streamlit dashboard. This application serves as the primary consumption layer, allowing users to explore exclusive pricing trends, alpha-generating factor indicators, and synthesized market sentiment. It provides an intuitive interface for research and signal monitoring.



Live Dashboard Demo

END