**Project Plan: Daily Market Data Analytics Pipeline**

**Status: Updated as of October 8, 2025**

**1. Executive Summary & Objectives**

**This project's goal is to design, build, and deploy a robust, automated data pipeline that ingests daily stock market data. The raw data will be processed and stored in a cloud data warehouse, where it will be modeled into an analytics-ready format. This foundational pipeline will enable historical market analysis and serve as a reliable data source for future business intelligence or data science initiatives.**

**Objectives:**

* **Objective 1: Automate Data Ingestion: Develop a reliable, scheduled process to extract daily stock market data.**
* **Objective 2: Create a Centralized Data Warehouse: Establish a clean, well-documented set of tables serving as the "single source of truth."**
* **Objective 3: Ensure Data Quality & Reliability: Implement automated checks and transformations to ensure data is accurate and consistent.**

**2. Technology Stack**

* **Cloud Provider: Google Cloud Platform (GCP) or Amazon Web Services (AWS)**
* **Data Ingestion: Python 3.x**
* **Data Source: Yahoo Finance (yfinance library) *(Decision made to simplify access and avoid API key management)***
* **Orchestration: Apache Airflow (running locally via Docker)**
* **Data Lake (Storage): Google Cloud Storage (GCS) or Amazon S3**
* **Data Warehouse: Google BigQuery or Snowflake**
* **Transformation & Modeling: dbt (Data Build Tool)**
* **Containerization: Docker & Docker Compose**

**3. Project Milestones & Timeline**

**Milestone 1: Environment Setup & Initial Extraction**

* **Status: ✅ COMPLETE**
* **Summary: The core Python extraction script has been successfully developed and tested. It correctly fetches data for our MVP ticker list and saves the output in a clean, partitioned folder structure. This work is formally accepted.**

**Milestone 2: Data Transformation & Modeling**

* **Status: ▶️ IN PROGRESS (Current Focus)**
* **Summary: The next critical step is to transform the raw JSON files into a structured, analytics-ready table.**
* **Tasks:**
  + **2.1: Initialize a dbt project and connect it to the data warehouse.**
  + **2.2: Configure dbt to read the raw JSON files from the local file system (and later, the data lake).**
  + **2.3: Build the dbt models (staging, dims, facts) to produce the final fct\_market\_daily table.**
  + **2.4: Implement data quality tests within dbt.**

**Milestone 3: End-to-End Pipeline Automation**

* **Status: ⏹️ UP NEXT**
* **Summary: This final phase will integrate all components into a fully automated, scheduled workflow using Apache Airflow.**
* **Tasks:**
  + **3.1: Modify the extraction script to upload the raw data directly to the GCS/S3 data lake.**
  + **3.2: Create an Airflow DAG with at least two tasks:**
    - **A task to run the Python extraction script.**
    - **A downstream task to execute dbt run after the extraction is successful.**
  + **3.3: Write a comprehensive README.md file for the project's GitHub repository.**