Orchestration Plan

Key Tasks

1. Ingest → acquire raw data.

2. Clean → preprocess and validate the data.

3. Train or Score → train or score on cleaned data.

4. Report → generate outputs for stakeholders.

Dependencies

Workflow: Ingest → Clean → Train or Score → Report.

Diagram:

```
{'ingest': [],
 'clean': ['ingest'],
 'train_or_score': ['clean'],
 'report': ['train_or_score']}
```

Task Specifications

Ingest

       Inputs: /data/raw.ext

       Outputs: prices_raw.json

       Idempotent: Yes

       Log: start/end, rows, source URI. Checkpoint: prices_raw.json

Clean

       Inputs: prices_raw.json

       Outputs: prices_clean.json

       Idempotent: Yes

       Log: start/end, rows in/out. Checkpoint: prices_clean.json.

Train or Score

     Inputs: prices_clean.json

     Outputs: model.json

     Idempotent: Yes (if seeded)

     Log: params, metrics. Checkpoint: model.json.

Report

     Inputs: model.json

     Outputs: report.txt

     Idempotent: Yes

     Log: artifact path. Checkpoint: report.txt.

Failure Points & Retry Policy

- Clean: fix manually.

- Model: retry with adjusted parameters if training fails.

Automation

Automate: For ingest and clean, we can automate because the raw stock data is very similar, and we can use the same to get and clean the data

Manual: For the train or score and report, we need to customize because different data has different inner patterns, which means that we need to design different models to train. Besides, different PM like different reports, so we also need to manually output the reports.