Stock Data Pipeline Reference Guide

1. Core Concepts

A data pipeline is the path data takes from being raw (as it comes from an API or file) to being cleaned and analyzed. For example: Yahoo Finance API → Raw CSV → Clean Parquet → Beta Analysis. Each step makes the data more structured and reliable.

2. Key Tools

■ DuckDB

DuckDB is an in-process analytical database (like SQLite but optimized for analytics). It’s ideal for large analytical queries, runs inside Python, and can query Parquet, CSV, or Pandas DataFrames directly.

■ Parquet

Parquet is a columnar file format used for analytics. It’s compressed, schema-aware, and compatible with Pandas, DuckDB, and Spark. It enables smaller storage, faster queries, and native integration with Databricks.

3. The Medallion Architecture

Layer	Purpose	Description	Example
■ Bronze	Raw Data	Unmodified source data; keep it safe.	Raw yfinance CSVs.
■ Silver	Cleaned Data	Fixed types, removed duplicates, validated.	Parquet files of daily prices.
■ Gold	Analytical Data	Aggregated metrics, KPIs, modeling inputs.	Betas, returns, portfolio summaries.

4. Folder Structure

stock_project/ → data/ (bronze, silver, gold), db/ (DuckDB file), scripts/ (fetch, clean, compute), and analytics/ (visualization notebooks). This structure mimics Databricks' Medallion model.

5. Example Flow

1■■ Bronze → Save raw API pulls as CSV.

2■■ Silver → Clean data and save as Parquet.

3■■ Gold → Compute metrics (e.g., betas) and store results in DuckDB.

6. When to Use Spark / Databricks

Use Spark or Databricks later when data exceeds memory, you need distributed ETL, or want real-time analytics. For now, DuckDB + Parquet + Pandas/Polars is optimal.

7. Learning / Growth Roadmap

Stage	Focus	Stack
Now	ETL + Analysis	Pandas / DuckDB / Parquet
Soon	Larger Data, Parallel Compute	PySpark (local)
Later	Production-Scale, Multi-User	Databricks Cloud