**📝 Data Collection Task — Trading Bot Project**

**🎯 Objective**

We need **historical OHLCV data (Open, High, Low, Close, Volume)** for multiple crypto symbols and timeframes. This data will feed into our backtester and strategy engine to generate trading signals.

**📌 Scope of Work**

1. **Symbols**  
   Collect for:
   * BTC/USDT
   * ETH/USDT
   * BNB/USDT
   * (later, we’ll expand to 10–20 symbols)
2. **Timeframes**
   * 1m, 5m, 15m, 1h, 4h, 12h, 1d  
     (at least last **1–2 years** for intraday, 3+ years for daily).
3. **Data Source**
   * Binance (preferred) → provides reliable OHLCV.
   * Use either:
     + **Binance API + Python script** (ccxt or binance library), OR
     + Manual export from Binance web charts (if API is blocked).
4. **Storage Format**
   * Save into data/ folder.
   * Folder structure:
   * data/
   * BTCUSDT/
   * 5m.csv
   * 15m.csv
   * 1h.csv
   * ...
   * ETHUSDT/
   * 5m.csv
   * 15m.csv
   * ...
   * Each CSV must have:
   * ts, open, high, low, close, volume
   * 2021-01-01 00:00:00, 29000, 29500, 28900, 29300, 120.5
   * Ensure **timestamps are UTC**.
5. **Compression**
   * Save files as .csv.gz (compressed). Pandas can read them directly → reduces storage by ~70%.
6. **Update Process**
   * Write/download script that can **append new candles daily** (so we don’t re-download everything).
   * Keep data up-to-date.

**✅ Deliverables**

1. Organized data/ folder with **symbols × timeframes** CSV files.
2. Each CSV has **consistent format** (ts, open, high, low, close, volume).
3. Compressed files (.csv.gz).
4. (Optional, bonus) A **Python script** to auto-download & update from Binance API.

**🔮 Why This Matters**

* Without reliable historical data, our backtests and signal generation are meaningless.
* Stage 2 and Stage 3 of our pipeline require enough **trade samples** (50–100+ trades) to calculate win probability, RR ratio, and confidence levels.
* More symbols + timeframes = richer strategy evaluation.

**Task Abstract: Automated Historical Crypto Data Collection**

**Objective:**  
Collect complete historical OHLCV (Open, High, Low, Close, Volume) data for selected crypto trading pairs (symbols) across multiple timeframes (e.g., 5m, 15m, 1h, 12h) for use in backtesting and live signal generation.

**Approach:**

1. Use a Python script to programmatically download CSV files from Binance’s official data repository.
2. Organize downloaded data into structured folders by symbol and timeframe (e.g., data/BTCUSDT/5m.csv).
3. The script should:
   * Check for existing files and skip duplicates.
   * Handle large datasets efficiently by fetching monthly or yearly chunks.
   * Merge multiple files into a single continuous dataset per symbol/timeframe.
   * Optionally prune older data beyond a retention period to save storage.
4. Ensure that the downloaded data is clean, correctly typed, and ready for backtesting scripts.

**Deliverable:**

* A Python script that can fetch, update, and maintain historical OHLCV CSV datasets automatically.
* A folder structure with all symbols and timeframes properly saved.
* A report/log confirming successful download and number of rows per file.