# Tổng kết Stage 1 — Data Foundation (Phiên bản Baseline Hoàn Chỉnh)

Tài liệu này tổng hợp toàn bộ các công việc đã hoàn thành trong Stage 1 của dự án AlphaForge — giai đoạn Xây dựng Nền Dữ Liệu. Bao gồm toàn bộ các bước từ 1.1 đến 1.11, nêu rõ mục tiêu, kết quả, và những file dữ liệu baseline được đóng băng để sử dụng cho Stage 2 (Feature Fusion).

## 1️⃣ Stage 1.1 → 1.7 — Dữ liệu thị trường OANDA

Mục tiêu: Thu thập và chuẩn hóa dữ liệu giá FX (OHLCV) từ OANDA. Các dữ liệu đều tuân thủ schema ['mid\_o', 'mid\_h', 'mid\_l', 'mid\_c', 'volume'] với alias 'close = mid\_c', và được lưu định dạng Parquet với index thời gian UTC.

✅ File baseline: data/GBP\_USD\_M5\_2024.parquet (và các cặp tương tự)

## 2️⃣ Stage 1.8 — Lịch kinh tế (Investing.com)

Mục tiêu: Thu thập toàn bộ lịch công bố dữ liệu kinh tế (2023-01-01 → 2025-10-30) từ Investing.com, bao gồm các trường thông tin event, impact, actual, forecast, previous và mức độ biến động (impact level).

✅ Script baseline: stage1\_8\_fetch\_investing\_calendar.py, stage1\_8\_qa\_audit\_calendar.py, stage1\_9\_normalize\_calendar.py

✅ Dữ liệu baseline:  
- data/econ\_calendar\_investing\_raw.parquet  
- data/econ\_calendar\_features.parquet

## 3️⃣ Stage 1.10 — Dữ liệu vĩ mô (Macro FX Drivers - FRED)

Mục tiêu: Thu thập các chỉ báo vĩ mô có ảnh hưởng mạnh đến thị trường ngoại hối. Bao gồm: UST2Y, UST5Y, UST10Y, UST10Y\_REAL, VIX, SPX, WTI, DE/UK/JP long-term rates, và các biến spread được tính toán tự động.

✅ Script baseline: stage1\_10\_fetch\_macro\_fxdrivers.py, stage1\_10\_qa\_audit\_macro\_fxdrivers.py, stage1\_10b\_rolling\_corr.py

✅ Dữ liệu baseline: data/macro\_fxdrivers.parquet

## 4️⃣ Stage 1.11 — Chỉ báo thanh khoản & nguồn vốn (Liquidity & Funding)

Mục tiêu: Xây dựng bộ dữ liệu đo lường thanh khoản và điều kiện nguồn vốn USD. Bao gồm WALCL (Fed BalanceSheet), RRP Usage, SOFR, EFFR, TBILL3M và các biến tính toán: Δ7d, spreads, chỉ số tổng hợp thanh khoản (LIQ\_COMPOSITE) và trung bình động 30 ngày.

✅ Script baseline: stage1\_11\_fetch\_liquidity.py, stage1\_11b\_qa\_liquidity.py

✅ Dữ liệu baseline: data/liquidity\_funding.parquet

## 🧩 Dữ liệu baseline cuối cùng cho Stage 2

Các dataset dưới đây được đóng băng (freeze) và sẽ dùng làm input cho Stage 2 — Feature Fusion & Alignment:  
- data/GBP\_USD\_M5\_2024.parquet (Giá OANDA)  
- data/econ\_calendar\_features.parquet (Sự kiện kinh tế)  
- data/macro\_fxdrivers.parquet (Chỉ báo vĩ mô)  
- data/liquidity\_funding.parquet (Thanh khoản & Funding)

## ⚙️ Các file dư thừa / không còn sử dụng

- stage1\_8a\_fix\_calendar\_mapping.py, stage1\_8b\_fix\_calendar\_mapping.py (đã được thay thế bởi crawler hoàn chỉnh)  
- econ\_calendar\_final.parquet, econ\_calendar\_investing.parquet (cũ, hợp nhất vào bản chính)  
- stage1\_10b\_rolling\_corr.py, stage1\_11b\_qa\_liquidity.py (chỉ dùng tham khảo, không nằm trong pipeline chính)

## 🧭 Trạng thái hoàn tất Stage 1

✅ Tất cả các bước trong Stage 1 đã hoàn thành, được kiểm định và đóng băng baseline.  
Toàn bộ dữ liệu chuẩn UTC, định dạng Parquet, không thiếu dữ liệu trọng yếu.  
Sẵn sàng chuyển sang Stage 2 — Feature Fusion & Time Alignment.