# AlphaForge — Stage 2: Data Integration & Feature Fusion (Tóm tắt)

Mục tiêu tổng quát:  
Stage 2 hợp nhất toàn bộ dữ liệu vĩ mô, thanh khoản và sự kiện kinh tế với dữ liệu giá FX M5 chuẩn (được chuẩn hoá từ Stage 1).  
Kết quả là một feature universe hoàn chỉnh – khung thời gian M5 thống nhất, sẵn sàng cho Feature Engineering và Machine Learning (Stage 3).

## 1. Các bước chính và kết quả đạt được

• 2.1 – Time Alignment Engine → data/fx\_timegrid.parquet  
• 2.2 – Merge FX Prices → data/stage2\_prices\_merged.parquet  
• 2.3 – Merge Macro Drivers → data/stage2\_macro\_merged.parquet  
• 2.4 – Merge Economic Calendar → data/stage2\_calendar\_merged.parquet  
• 2.5 – Merge Liquidity & Funding → data/stage2\_liquidity\_merged.parquet  
• 2.6 – Unified Feature Fusion → data/stage2\_features\_combined.parquet

## 2. Kết quả tổng thể

• Tất cả dữ liệu được chuẩn hóa theo UTC, tần suất 5 phút, index DatetimeIndex duy nhất.  
• Không còn lỗi NaN lớn, không lệch thời gian.  
• Các prefix chuẩn hóa:  
 - px\_ = giá FX  
 - macro\_ = chỉ báo vĩ mô  
 - cal\_ = sự kiện kinh tế  
 - liq\_ = chỉ báo thanh khoản

## 3. Dataset baseline Stage 2 (đã freeze)

• data/fx\_timegrid.parquet  
• data/stage2\_prices\_merged.parquet  
• data/stage2\_macro\_merged.parquet  
• data/stage2\_calendar\_merged.parquet  
• data/stage2\_liquidity\_merged.parquet  
• data/stage2\_features\_combined.parquet

## 4. Kết luận

Stage 2 hoàn tất thành công.  
Từ Stage 3 trở đi, toàn bộ hệ thống ML sẽ sử dụng duy nhất các baseline sau:  
- data/stage2\_features\_combined.parquet (master input)  
- data/fx\_timegrid.parquet (time reference)  
Các file khác được coi là source frozen, không chỉnh sửa nữa.

## 5. Tiếp theo — Stage 3: Feature Engineering & Labelling

• Tạo nhãn y (±1 / 0 = Up / Flat / Down)  
• Sinh feature tsfresh & technical indicators  
• Chuẩn hóa dữ liệu cho LightGBM training