Cryptocurrency Data Pipeline

Project Report

Team Members

Student ID	Name
22B031332	Shim Daniil
22B030340	Jakupov Dias
22B030372	Kabiyev Zhanbek

Project Overview

This project implements a real-time cryptocurrency market data pipeline using Apache Airflow, Apache Kafka, and SQLite. The system collects global market metrics from the CoinMarketCap API, processes the data through a streaming pipeline, and generates daily analytics with visualizations.

Technology Stack:

- Apache Airflow 2.7.0 - Workflow orchestration

- Apache Kafka - Message streaming

- PostgreSQL - Airflow metadata database

- SQLite - Application data storage

- Docker Compose - Container orchestration

- Python (pandas, matplotlib) - Data processing and visualization

Job 1: Data Ingestion (Producer)

Purpose: Fetches real-time cryptocurrency market data from CoinMarketCap API and publishes to Kafka.

Schedule: Runs every 5 minutes via Airflow DAG

Process:

- Connects to CoinMarketCap Global Metrics API
- Fetches data every 30 seconds for 5 minutes per DAG run
- Publishes raw JSON responses to Kafka topic 'raw_events'
- Handles API errors gracefully with retry logic

Data Collected:

- Total market cap and 24h volume
- BTC and ETH dominance percentages
- DeFi, Stablecoin, and Derivatives metrics
- Active cryptocurrencies and exchanges count

Job 2: Data Cleaning and Storage

Purpose: Consumes raw data from Kafka, cleans it, and stores in SQLite database.

Schedule: Runs every hour via Airflow DAG

Cleaning Operations:

- Parses nested JSON structure from CoinMarketCap API
- Converts timestamps to datetime format
- Removes records with null timestamp or market cap
- Deduplicates records based on timestamp
- Clips negative values in numeric columns to 0
- Forward-fills missing dominance values

Output: Cleaned records stored in 'events' table with 17 metrics per record

Job 3: Daily Analytics and Visualization

Purpose: Computes daily aggregations and generates visualization charts.

Schedule: Runs daily at midnight via Airflow DAG

Analytics Computed:

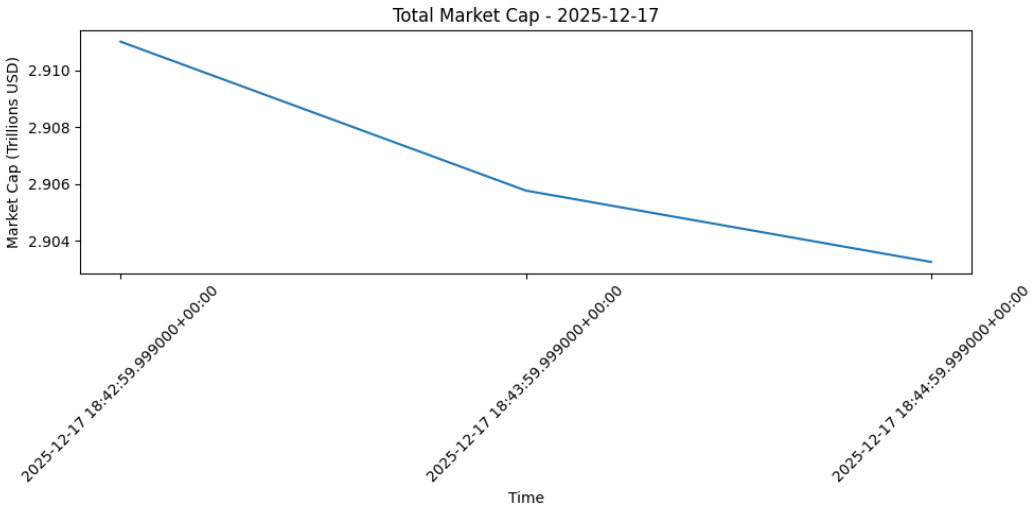
- Average, min, max market cap and volume
- BTC/ETH dominance statistics (avg, min, max, std dev, percentiles)
- Daily BTC dominance change
- DeFi, Stablecoin, and Derivatives averages
- High volatility event count (>5% daily change)
- Records per hour metric

Visualizations Generated:

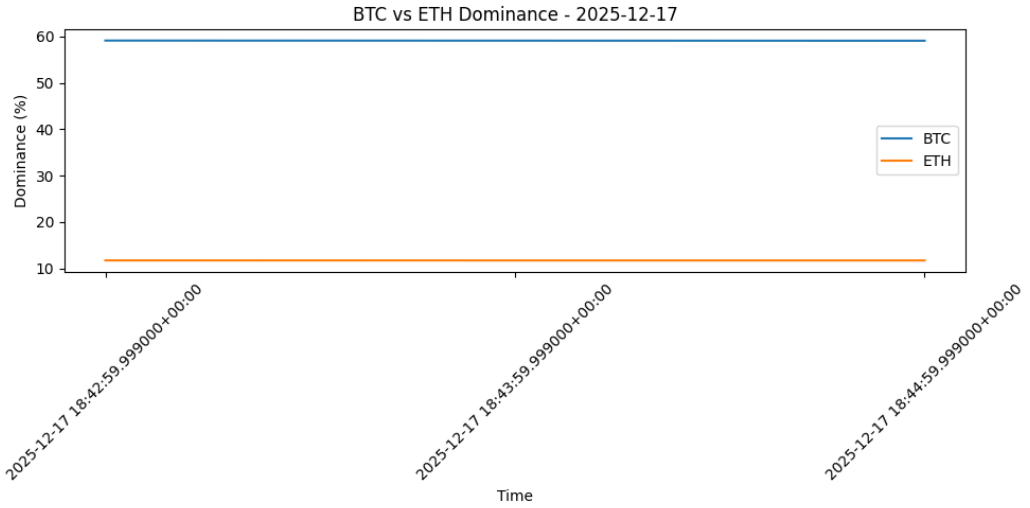
- Market Cap Timeline - Line chart showing intraday market cap movement
- BTC vs ETH Dominance - Dual line chart comparing dominance trends
- Market Segments Pie - Distribution of DeFi, Stablecoins, and Altcoins

Generated Charts

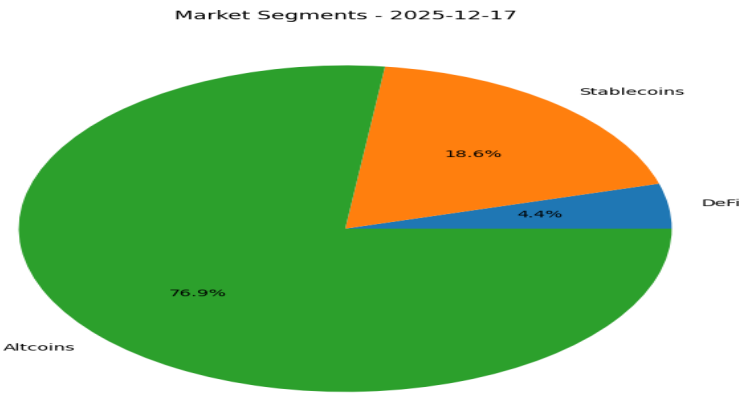
Total Market Cap Timeline



BTC vs ETH Dominance

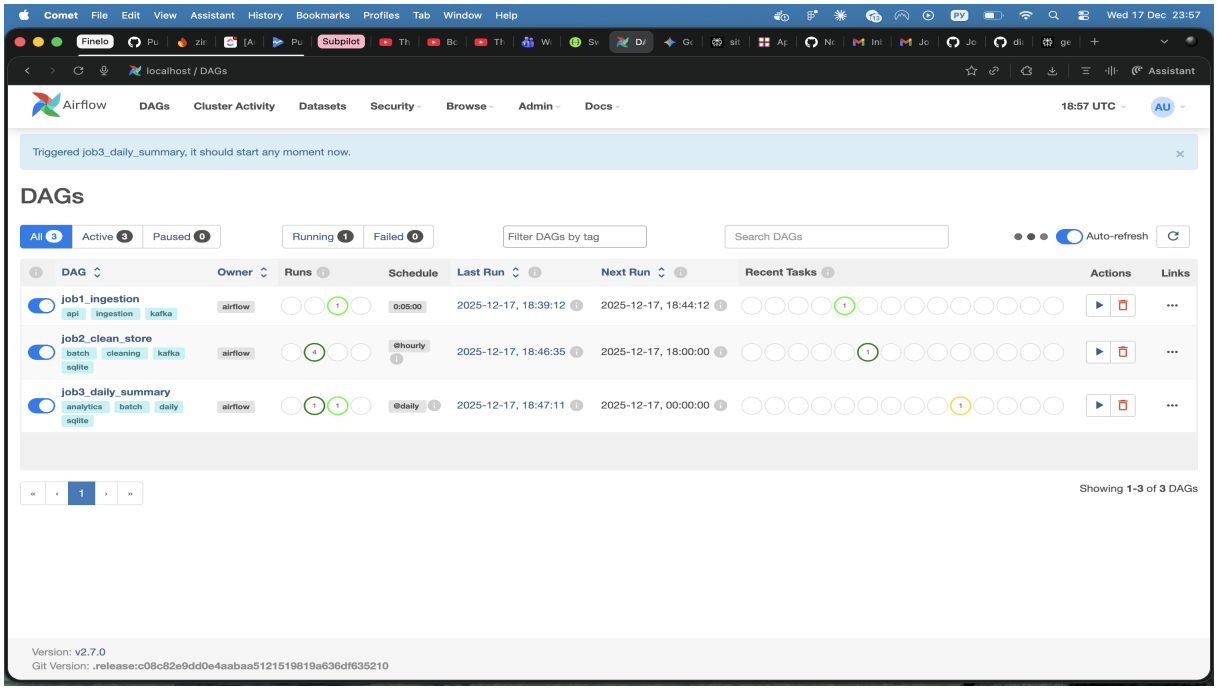


Market Segments Distribution

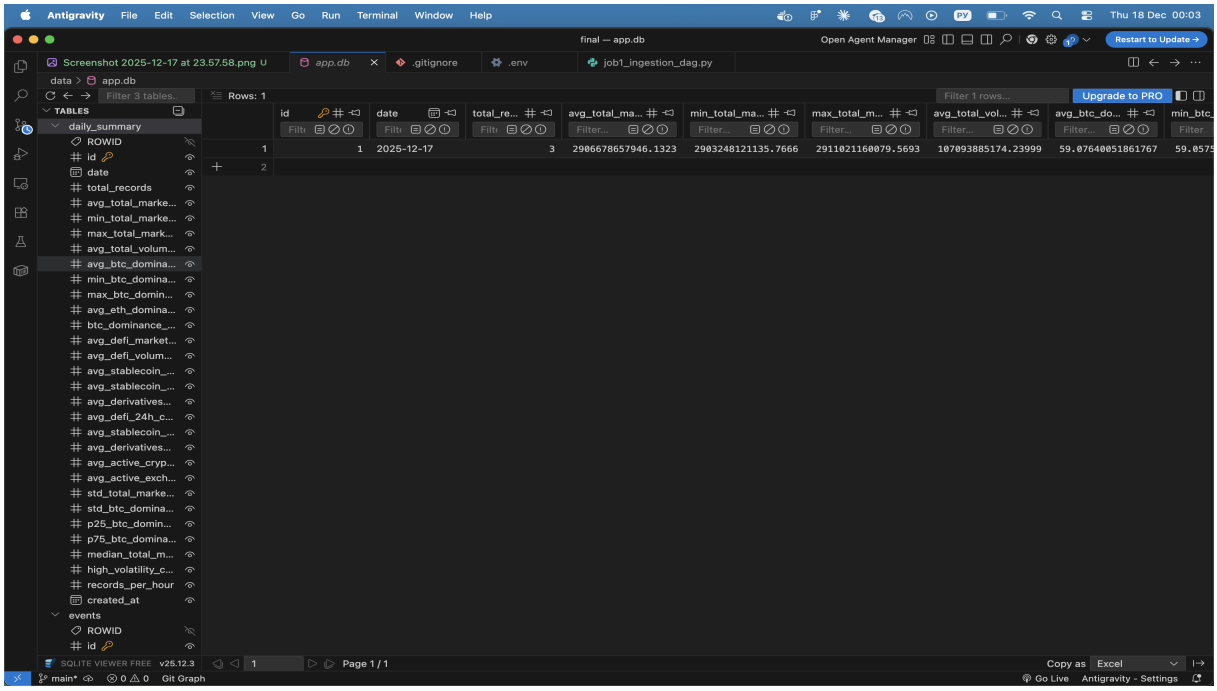

Market Segments - 2025-12-17
Stablecoins
18.6%
DeFi
4.4%
76.9%
Altcoins

System Screenshots

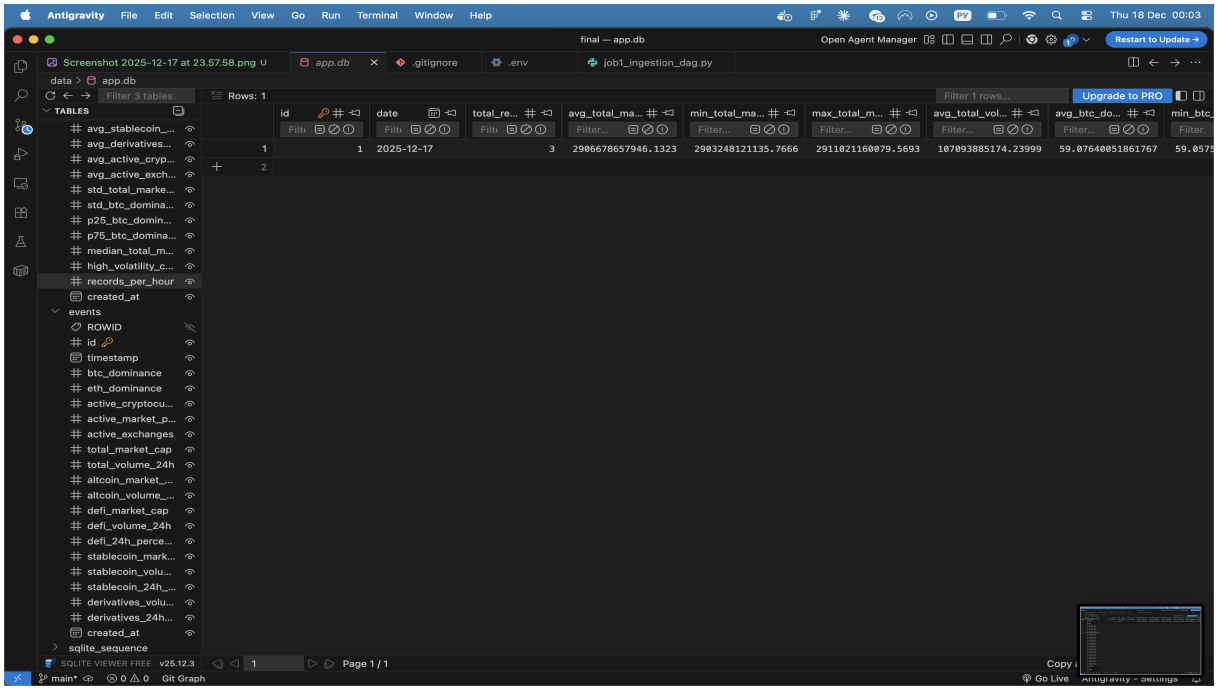
Airflow DAGs Dashboard - Shows all three pipeline jobs



Daily Summary Table - Aggregated analytics data



Daily Summary Schema - Database structure



Events Table - Raw cleaned data records