

# 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:**

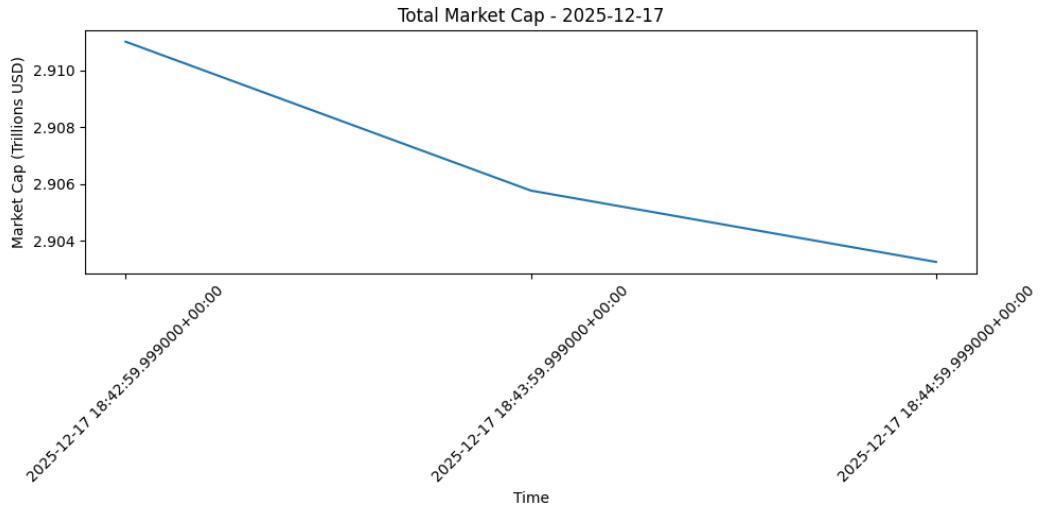
- Average, min, max market cap and volume
- BTC/ETH dominance statistics (avg, min, max, std dev, percentiles)
- Daily BTC dominance change
- DeFi, Stablecoin, and Derivatives averages
- High volatility event count (>5% daily change)
- Records per hour metric

### **Visualizations Generated:**

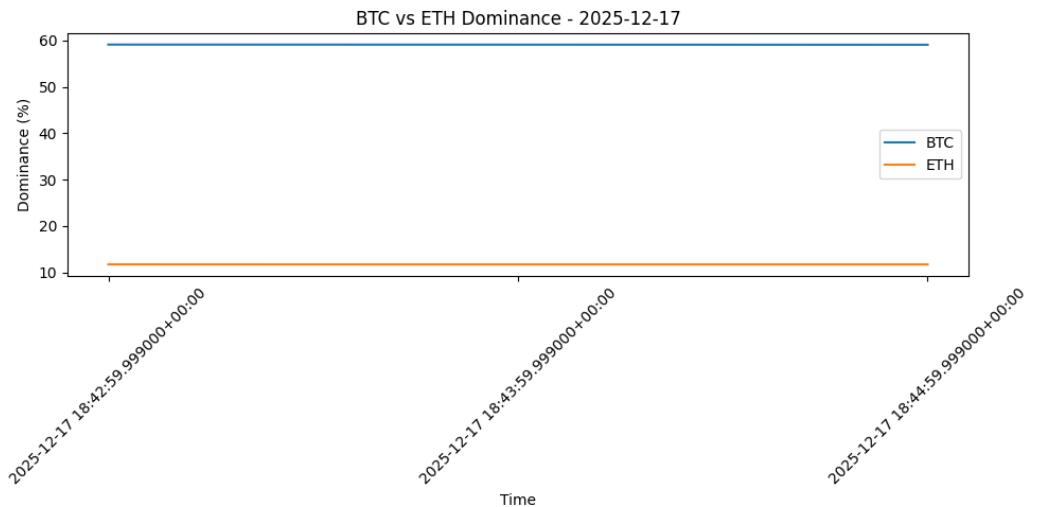
- Market Cap Timeline - Line chart showing intraday market cap movement
- BTC vs ETH Dominance - Dual line chart comparing dominance trends
- Market Segments Pie - Distribution of DeFi, Stablecoins, and Altcoins

# Generated Charts

## Total Market Cap Timeline

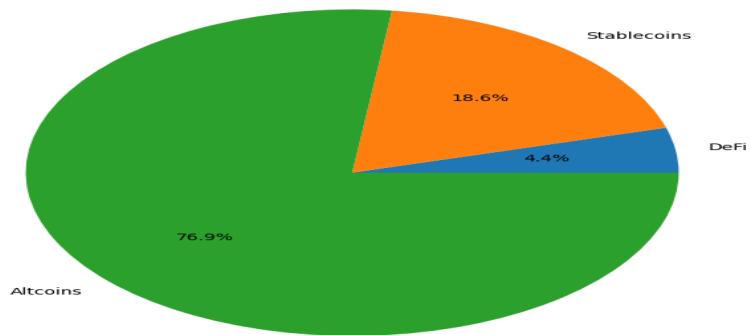


## BTC vs ETH Dominance



## Market Segments Distribution

Market Segments - 2025-12-17



# System Screenshots

## Airflow DAGs Dashboard - Shows all three pipeline jobs

The screenshot shows the Airflow DAGs dashboard with three pipeline jobs listed:

DAG	Owner	Runs	Schedule	Last Run	Next Run	Recent Tasks	Actions	Links
job1_ingestion	airflow	1	0:05:00	2025-12-17, 18:39:12	2025-12-17, 18:44:12	1	[Run]	[View]
job2_clean_store	airflow	4	@hourly	2025-12-17, 18:46:35	2025-12-17, 18:00:00	1	[Run]	[View]
job3_daily_summary	airflow	1	@daily	2025-12-17, 18:47:11	2025-12-17, 00:00:00	1	[Run]	[View]

Version: v2.7.0  
Git Version: release:c08c82e9dd0e4aabaa5121519819a636df635210

## Daily Summary Table - Aggregated analytics data

The screenshot shows the Antigravity SQLite viewer interface. The title bar indicates the file is named 'final — app.db'. The main window displays the 'daily\_summary' table with two rows of data. The columns represent various aggregated metrics such as total records, average total market value, and maximum total market value. The table has 11 columns: id, date, total\_records, avg\_total\_marketing\_value, min\_total\_marketing\_value, max\_total\_marketing\_value, avg\_total\_volume, avg\_btc\_dominance, min\_btc\_dominance, avg\_defi\_market\_value, and avg\_stablecoin\_value.

id	date	total_records	avg_total_marketing_value	min_total_marketing_value	max_total_marketing_value	avg_total_volume	avg_btc_dominance	min_btc_dominance	avg_defi_market_value	avg_stablecoin_value
1	2025-12-17	3	2906678657946.1323	2903248121135.7666	2911021160079.5693	107093885174.23999	59.07640085186176	59.0575		
2										

Below the table, there is a detailed list of all columns in the 'daily\_summary' table, including their names and descriptions. The interface also shows other tables like 'events' and 'ROWID'.

## Daily Summary Schema - Database structure

The screenshot shows the Antigravity database viewer interface. The title bar indicates the file is 'final -- app.db'. The main window displays the schema and data for the 'app.db' database. The schema tree on the left shows various tables such as 'avg\_stablecoin...', 'avg\_DERIVATIVES...', 'avg\_ACTIVE\_CRYPT...', 'std\_total\_mark...', 'std\_btc\_domin...', 'p25\_btc\_domin...', 'p75\_btc\_domin...', 'median\_total\_m...', 'high\_volatility\_c...', 'records\_per\_hour...', 'events', 'timestamp', 'btc\_dominance', 'eth\_dominance', 'active\_crypto...', 'active\_market...', 'active\_exchanges', 'total\_market\_cap', 'total\_volume\_24h', 'altcoin\_market...', 'altcoin\_volume...', 'defi\_market\_cap', 'defi\_volume\_24h', 'defi\_24h\_percent...', 'stablecoin\_mark...', 'stablecoin\_volum...', 'stablecoin\_24h...', 'derivatives\_volum...', 'derivatives\_24h...', and 'sqlite\_sequence'. The data table on the right shows one row of data with columns: id, date, total\_re..., avg\_total\_ma..., min\_total\_ma..., max\_total\_ma..., avg\_total\_volum..., avg\_btc\_domin..., and min\_btc\_domin....

id	date	total_re...	avg_total_ma...	min_total_ma...	max_total_ma...	avg_total_volum...	avg_btc_domin...	min_btc_domin...
1	2025-12-17	3	2906678657946.1323	2903248121135.7666	2911021160079.5693	107093885174.23999	59.07640085186176	59.0575

## Events Table - Raw cleaned data records

The screenshot shows the Antigravity SQLite viewer interface. The title bar reads "Antigravity" and "Thu 18 Dec 00:03". The main window displays the "events" table from a database named "final — app.db". The table has 10 columns: id, timestamp, btc\_dominance, eth\_dominance, active\_cryptocurrencies, active\_markets, active\_exchanges, total\_market\_cap, total\_volume\_24h, and defl\_24h\_percent. There are three rows of data:

	1	2025-12-17 18:42:59.999000+000	59.095789217998	11.742658061168	9002	117064	907	2911021160079,5693	107
	2	2025-12-17 18:43:59.999000+000	59.075985923826	11.735568311558	9002	117064	907	2985766692623,061	106
	3	2025-12-17 18:44:59.999000+000	59.057586414929	11.734888607588	9002	117064	907	2983248121135,7666	107