Crypto-Media Integration: A NoSQL Data Pipeline for Market Sentiment Analysis

Data Management Project Report

Tommaso Arone
ID: 896282

Lorenzo Zanotti
ID: 902652

Lorenzo Triolo
ID: 895541

January 11, 2026

Agenda

The Context:

- Cryptocurrency markets are highly reactive to external information (regulations, adoption).
- Information flow is hard to measure directly.

The Project Goal:

- Use **The New York Times** coverage as a proxy for mainstream media attention.
- Analyze impact on **Bitcoin (BTC)** and **Ethereum (ETH)**.
- Construct a full NoSQL data pipeline (Ingestion $\rightarrow$ Storage $\rightarrow$ Analytics).

The study is guided by two main questions:

RQ1: Association

Is there a statistical association between NYT news volume about cryptocurrencies and Bitcoin/Ethereum price/volume?

RQ2: Predictive Power

Does media attention behave as a *leading* indicator for market movements (e.g., next-day or next-week changes)?

Data Sources & Ingestion Strategy

1. **Financial Data (Structured)**
   - **Source:** Yahoo Finance API ('yfinance').
   - **Assets:** BTC-USD, ETH-USD.
   - **Process:** ETL pipeline, reshaping "Wide" to "Long" format, handling 'NaN' values.
2. **News Media Data (Unstructured)**
   - **Source:** NYT Article Search API.
   - **Keywords:** "Bitcoin", "Ethereum", "Crypto".
   - **Constraints:** Strict rate limits (HTTP 429 handling).
   - **Mechanism:** Decoupled Producer-Consumer via **Apache Kafka** to ensure fault tolerance.

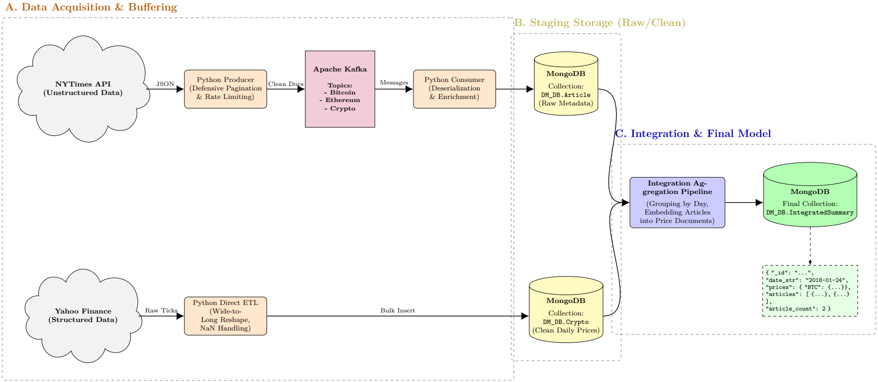


Figure: Producer-Consumer Architecture with Kafka and MongoDB.

Why MongoDB?

We selected a NoSQL Document Store (**MongoDB**) for three reasons:

1. **Semi-Structured Data:** Handles polymorphic news metadata (varying abstract lengths, missing fields).
2. **Data Locality (Embedding):** Articles are embedded *inside* the daily market document. Eliminates expensive `JOIN`s during analysis.
3. **Time-Based Bucketing:** The "Daily Bucket" model aggregates all events (News + Price) for a single day t_0.

Integration Logic: Context & Trends

The Pipeline creates a **Materialized View** (daily_market_summary) storing calculated future trends (t_{+1}, t_{+7}) in the current document (t_0).

```
{
  "_id": {
    "$oid": "694588d1d65f26c5a42ef2f8"
  },
  "date": {
    "$date": "2018-01-24T00:00:00.000Z"
  },
  "date_str": "2018-01-24",
  "articles": [
    {
      "remote_id": "...",
      "title": "...",
      "abstract": "..."
    }
  ],
  "article_count": 2,
  "prices": {
    "BTC-USD": {
      "open": 10903.40,
      "close": 11359.40,
      "intraday_change": 456,
      "context": { ... },
      "future_trend": { ... }
    }
    "ETH-USD": {...}
  }
}
```

```
"context": {
  "close_prev_day": 10868.400390625,
  "close_next_day": 11259.400390625,
  "close_prev_week": 11188.599609375,
  "close_next_week": 10221.099609375
}
```

```
"future_trend": {
  "next_day_diff": -100,
  "next_week_diff": -1138.3
}
```

High-Frequency Data Management Tasks

The following queries serve as benchmarks for evaluating the database's performance.

They demonstrate the schema's ability to seamlessly correlate:

- Unstructured news metadata (NYT).
- Time-series financial data (Bitcoin/Ethereum).

Goal: Show how complex analytical questions can be answered with single aggregation pipelines, avoiding expensive client-side processing.

Q1: Implementation

Goal: Find days with the most news articles.

```
pipeline_1 = [
    # 1. Sort for article_count (descending) to find peaks
    { "$sort": { "article_count": -1 } },
    { "$limit": 5 },

    # 2. Project: Reshape data to correlate News Volume with Price Action
    {
        "$project": {
            "date": "$date_str",
            "#articles": "$article_count",
            "title": { "$arrayElemAt": ["$articles.title", 0] },
            # Embedding Financial Context directly in the output
            "btc_close": "$prices.BTC-USD.close",
            "btc_vol": "$prices.BTC-USD.intraday_change",
            "eth_vol": "$prices.ETH-USD.intraday_change",
        }
    }
]
```

Q1: Results (News Peaks vs Volatility)

Finding: High volume correlates with major market shifts (e.g., FTX collapse).

Date	#Arts	Top News Title	BTC Close	BTC Vol.	ETH Vol.
2022-12-13	15	In FTX Collapse...	17,781	+574.88	+45.89
2022-11-30	13	Before FTX fell...	17,168	+723.09	+78.76
2018-06-28	13	A Field Guide...	5,903	-249.72	-19.93
2021-05-13	12	Elon Musk Makes...	49,716	-19.24	-113.77
2021-04-14	12	Coinbase Listing...	63,109	-414.06	+135.76

Q2: Implementation

Goal: Classify recovery types ("V-Shape" vs "U-Shape") after a crash.

```
pipeline_2 = [
    {   # 1. Filter: Focus only on significant BTC drops (< -50)
        "$match": { "prices.BTC-USD.intraday_change": { "$lt": -50 } }
    },
    {
        "$project": {
            "Date": "$date_str", "Articles": "$article_count",
            "BTC_Crash": { "$subtract": ["$prices.BTC-USD.close", "$prices.BTC-USD.context.close_prev_week"] },

            # 2. Logic: Classify Resilience based on future trend (t+1, t+7)
            "BTC_Recovery_Type": {
                "$switch": {
                    "branches": [
                        # If price rebounds next day -> V-Shape
                        { "case": { "$gt": ["$prices.BTC-USD.future_trend.next_day_diff", 0] }, "then": "V-Shape (Fast)" },
                        # If price rebounds next week -> U-Shape
                        { "case": { "$gt": ["$prices.BTC-USD.future_trend.next_week_diff", 0] }, "then": "U-Shape (Slow)" }
                    ],
                    "default": "No Recovery"
                }
            }
        }
    },
    { "$sort": { "BTC_Crash": 1 } }, { "$limit": 5 }
]
```

Q2: Results (Recovery Classification)

Objective: Analyze market resilience.

- **V-Shape:** Rapid 24h rebound.
- **U-Shape:** Slower weekly recovery.

Date	Arts	BTC Crash	BTC 24h	BTC 7d	BTC Rec.	ETH Rec.
2025-11-17	11	-13,902	+855	-3,823	V-Shape	V-Shape
2021-05-18	4	-13,795	-5,906	-4,507	No Rec.	No Rec.
2025-03-09	1	-13,647	-2,069	+1,978	U-Shape	No Rec.
2024-08-05	2	-12,828	+2,042	+5,363	V-Shape	V-Shape

Goal: Statistical baseline comparing days WITH vs WITHOUT news.

```
pipeline_3 = [
    {
        "$group": {
            # 1. Conditional logic to create dynamic groups
            "_id": {
                "$cond": [{ "$gt": ["$article_count", 0] }, "Days WITH News", "Days
     WITHOUT News"]
            },
            "total_days": { "$sum": 1 },

            # 2. Calculate average ABSOLUTE volatility (magnitude of move)
            "btc_avg_abs_volatility": { "$avg": { "$abs": "$prices.BTC-USD.
     intraday_change" } },
            "eth_avg_abs_volatility": { "$avg": { "$abs": "$prices.ETH-USD.
     intraday_change" } }
        }
    },
    { "$project": { "_id": 0, "Category": "$_id", "total_days": 1, "
     btc_avg_abs_volatility": 1, "eth_avg_abs_volatility": 1 } }
]
```

Q3: Results (Macro Volatility Baseline)

Finding: Volatility is $\approx 2x$ higher on days with media coverage.

Category	Total Days	BTC Avg Abs Vol	ETH Avg Abs Vol
Days WITH News	1632	$982.91	$65.52
Days WITHOUT News	1270	$474.03	$28.65

Quality Metrics Overview

1. **Duplication Rate**
   - Initial API fetch: 7,257 docs.
   - Duplicates: 27.7% (Multi-query overlap).
   - Deduplication Technique.
   - **Final:** 3,665 unique articles.
2. **Completeness**
   - Abstracts missing: 4.18%.
   - **Imputation:** ‘Abstract‘ $\leftarrow$ ‘Title‘.
   - Result: 0% missing textual data.
3. **Blind Spots (No News)**
   - Days with Market Data but **0** News.
   - **Result:** 43.76% of days are "Blind Spots".

Temporal Continuity

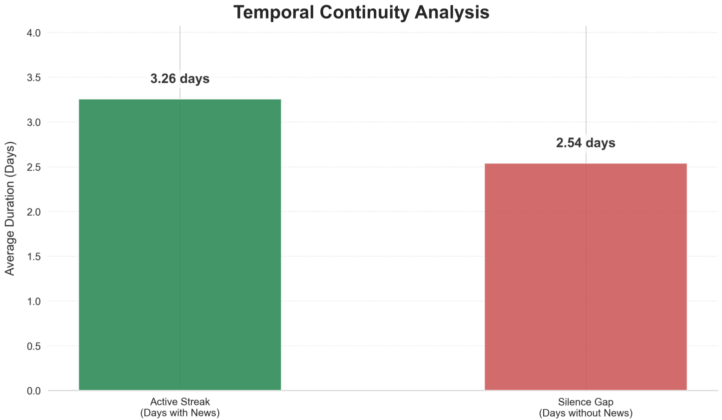


Figure: Information flow is fragmented: Average news streak is 3.26 days.

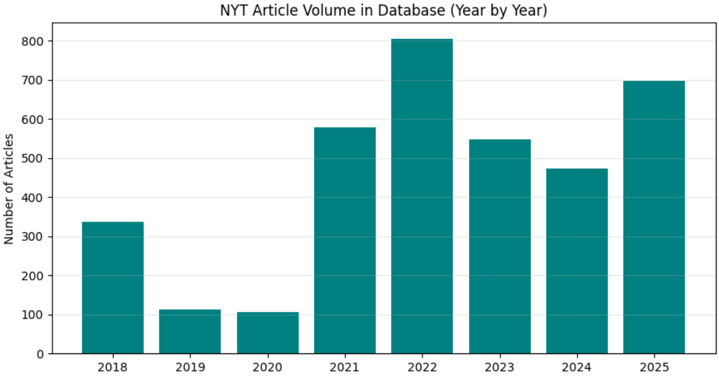
Semantic Redundancy

Question: On days with many articles, do they say the same thing?

- **Metric:** Pairwise Cosine Similarity (TF-IDF).
- **Result:** Global redundancy is low (2.14%).
- **Outliers:** High redundancy (> 0.20) occurs only during massive shocks and main events (e.g., Presidential Election).

Figure: Word clouds for high-redundancy events.

Temporal Distribution (Volume)



Key Insights from Report:

- **Regime Change (2021):** Structural break from "Crypto Winter".
- **Crisis Peak:** Max volume ($>$ 800 in 2022).
- **Event-Driven:** Media reacts to market collapses rather than tech adoption.

News Volume vs. Market Trends

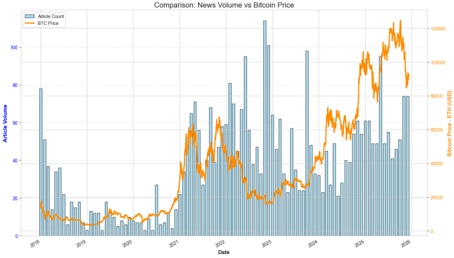


Figure: Bitcoin Price vs. Volume

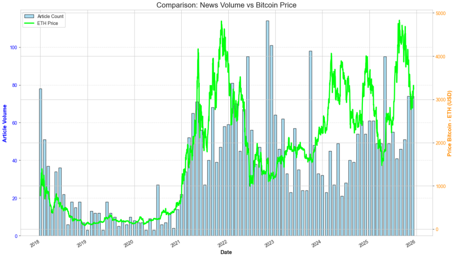
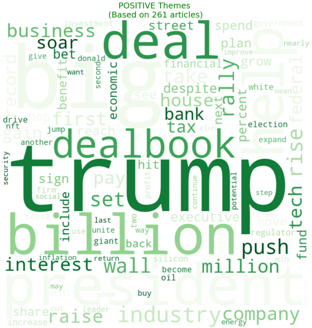


Figure: Ethereum Price vs. Volume

Core Findings:

- **Volatility Proxy:** Volume is a strong indicator of market stress (liquidity crises).
- **Asymmetric Coverage:** Crashes generate disproportionate spikes compared to rallies. Media acts as a *lagging* or *coincident* indicator.

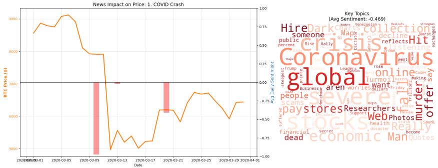


Narrative Dichotomy:

- **Positive Themes:** Linked to "Corporate Adoption" and "Growth" (Keywords: *Tech, Rally, Investment*).
- **Negative Themes:** Dominated by "Legal" and "Criminal" terminology rather than asset performance (Keywords: *Fraud, Federal, Prosecutor, Prison*).

Impact Analysis: Case Studies

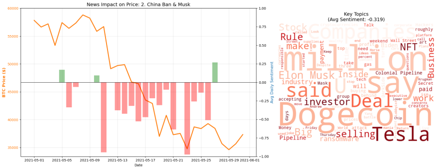
Case 1: The COVID Crash (Mar 2020)



External Shock (Macro)

- **Sentiment Collapse:** Score plummeted to **-0.469**.
- **Insight:** Terms like "Crisis" and "Global" confirm Crypto behaved as a correlated risk asset (high-beta) amid systemic panic.

Case 2: The China Ban (May 2021)



Industry Specific (Endogenous)

- **Specific Catalysts:** Volatility driven by identifiable actors ("Musk", "Tesla") and regulation ("China").
- **Insight:** Unlike Covid, this correction was narrative-driven with mixed sentiment signals.

Keyword Analysis: The "Trump" Factor

Keywords: President Trump

Political Framing:

- **Geopolitical Context:** Crypto is discussed through the lens of US Economic Policy (*"Tariff"*, *"Trade"*, *"Election"*).
- **Observation:** The narrative focuses on macro-economics rather than specific blockchain regulation.

Conclusions

Summary of Contributions & Key Findings:

- **Robust NoSQL Pipeline:** Decoupled Kafka-MongoDB architecture successfully correlated heterogeneous data (NYT + Crypto prices), enabling complex queries without expensive joins.
- **Volatility Correlation:** News volume acts as a stress indicator. Mainstream media is often *event-driven* (lagging/coincident), spiking during market failures (e.g., FTX).
- **Sentiment Dichotomy:** Positive coverage links to "Growth/Tech"; negative coverage is dominated by "Legal/Fraud" rather than price action.
- **Data Quality:** High information entropy when news is present, despite a 43% "Blind Spot" rate (days without coverage).

Future work

Future Work:

- **Advanced Sentiment Scoring:** Upgrade from keyword-based to Transformer models (BERT/FinBERT) for quantitative sentiment signals.
- **Predictive Modeling:** Use news features in ARIMAX/VAR models to forecast next-day volatility ($t + 1$).
- **Source Diversification:** mitigate "Blind Spots" by adding crypto-native media or social data (X/Twitter) to compare bias and latency.

Thank You