Data Engineering Project-

How Does Bitcoin News Affect Its Price?

Shaked Shtauber & Adi Segev

Naya College 06/24/2024

Table of Contents

Introduction

Project Overview

The objective of this project is to analyze how the news sentiment surrounding Bitcoin affects its market price. The core of the project lies in extracting news articles and price metrics on Bitcoin and then determining how these sentiments correlate with Bitcoin's price movements.

Motivation

Given the volatile nature of Bitcoin's price, understanding external factors that influence market fluctuations is crucial. In this case, Bitcoin-related news sentiment (positive, negative, or neutral) can provide valuable insights into price trends.

Objective

The goal is to build a pipeline that automates the collection and processing of Bitcoin news data and its price metrics and then correlates the sentiments with Bitcoin’s price movements.

Data Sources

- **Bitcoin Price Data**: Google Finance API is an API that allows us to connect to data from Google's financial page. This page provides information on stock values in various markets, as well as the value of cryptocurrencies.

  The data is categorized into arrays as follows:

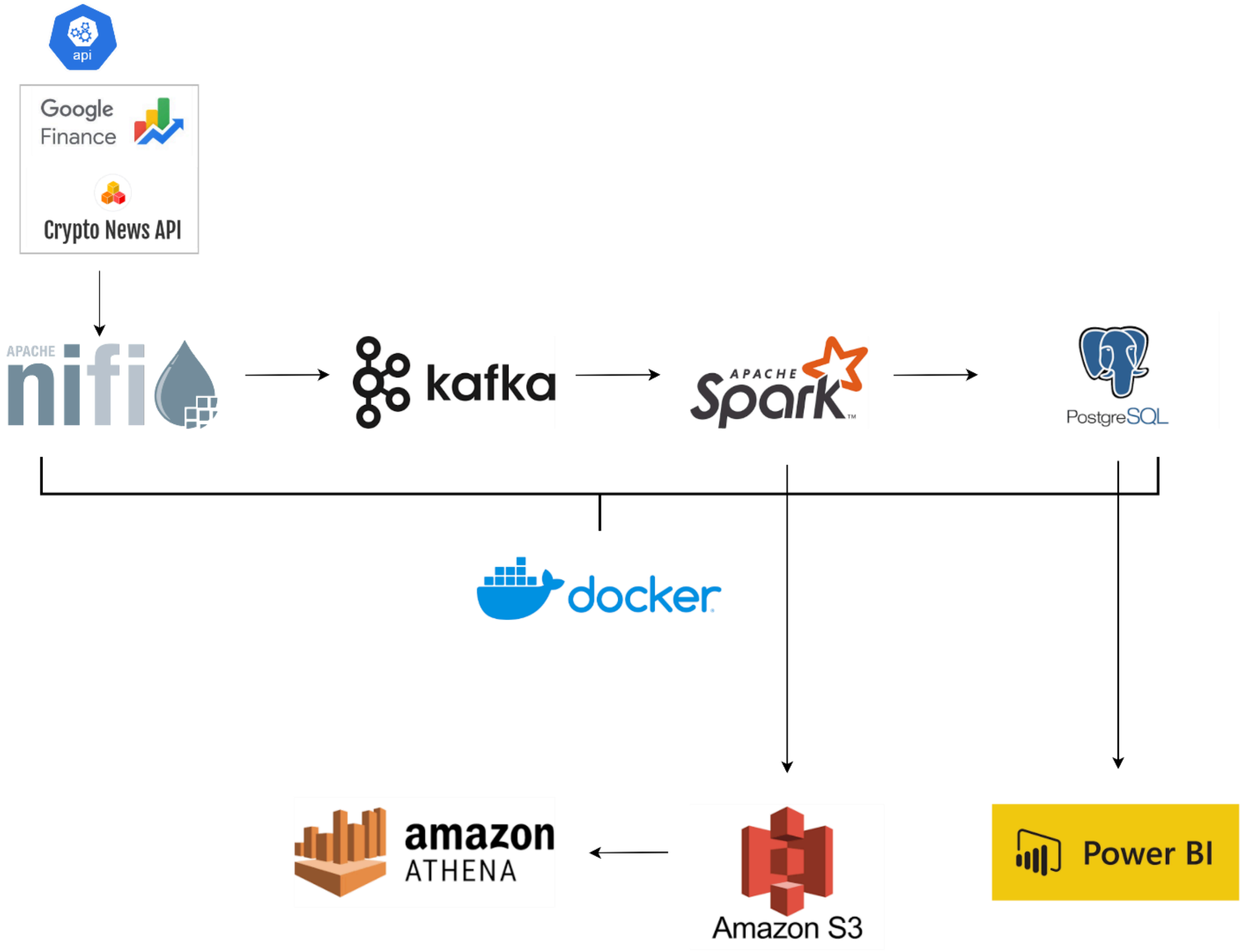
  - US Stocks
  - European Stocks
  - Asian Stocks
  - Cryptocurrency Stocks

- **Bitcoin News Articles**: We retrieve Bitcoin news data from Crypto News API**,** this contains the article sentiment. The sentiment classifies news articles into three categories:

  - **Positive Sentiment**: Articles that express optimism or favorable views about Bitcoin
  - **Negative Sentiment**: Articles that express caution, fear, or negative perspectives about Bitcoin
  - **Neutral Sentiment**: Articles that are impartial or present a balanced view

Data Engineering Process

Architecture:

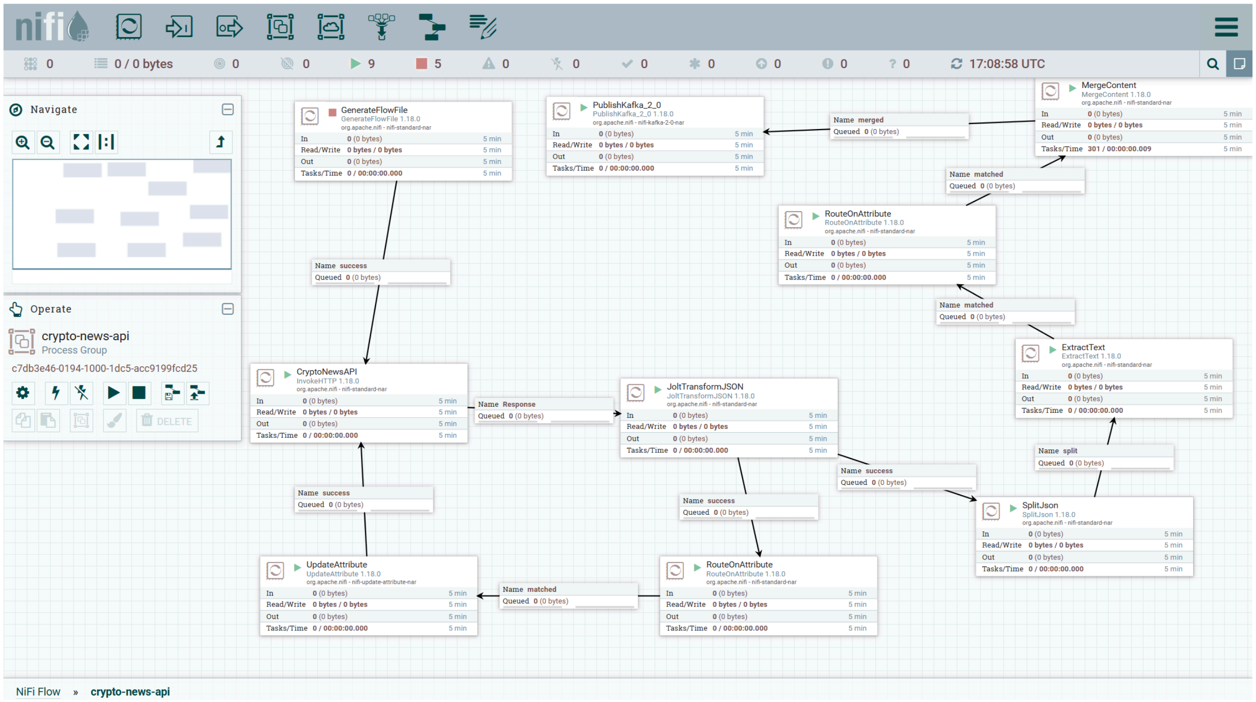


Data Pipeline:

1. **NiFi** - Retrieve the API and perform necessary changes to get our desired data in the right format.

   Crypto News API: We loop over 5 pages of records each containing 100 records which is the most back we can get with the basic subscription plan (we are only able to query up to 500 historical records at a time, 100 in a page). Our goal for doing that is to retrieve all the data we can get and filter for yesterday’s data. Then we choose the attributes we want followed by a split into individual Json records so we can filter on the date field. After the date comparison is done, we merge it back into one file and send it to Kafka. This flow is scheduled to run every morning at 6am. [we have data from 31/01/2025].

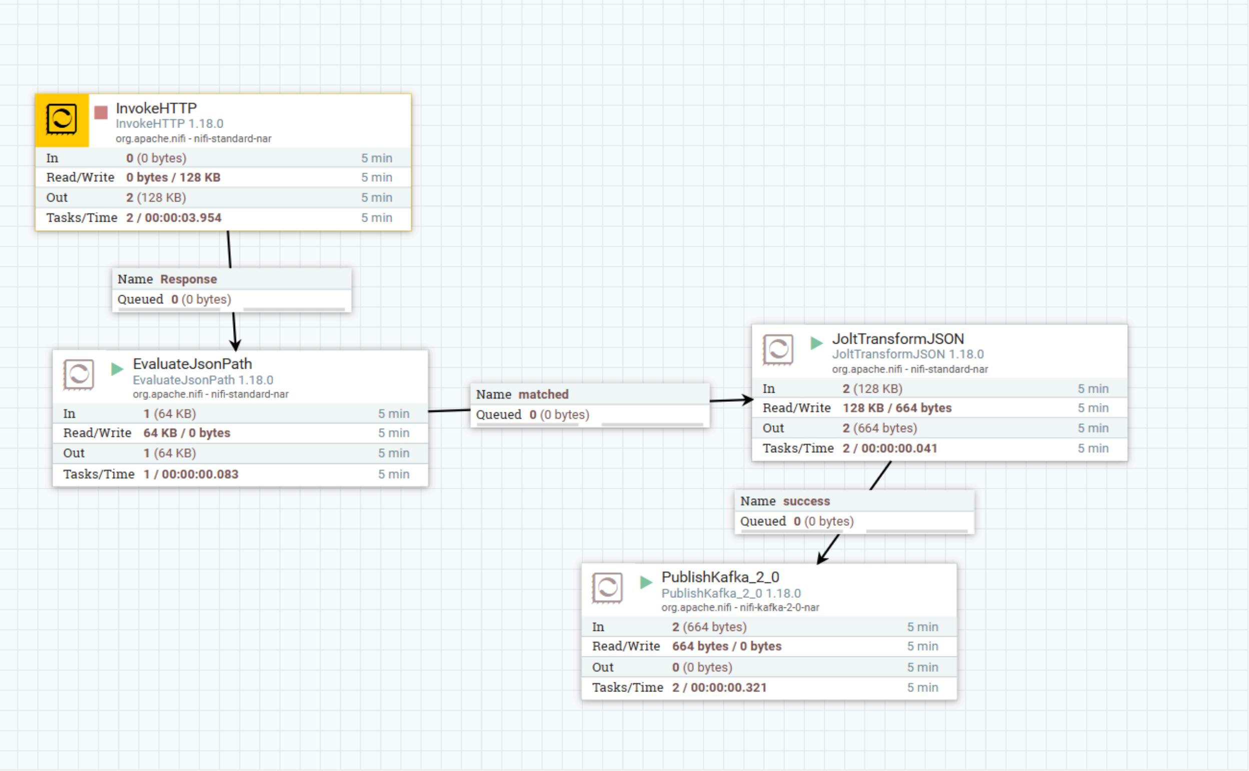
Crypto New API flow:



Google Finance API:

We specifically extract the financial values of Bitcoin from the cryptocurrency array, along with the date field from the general information array. The API has a built-in parameter that allows us to pull data from the last day, the last week, or the last month. For cost purposes, we use NIFI to pull the Bitcoin price data every hour.

Google Finance API flow:



InvokeHTTP
InvokeHTTP 1.18.0
org.apache.nifi - nifi-standard-nar

In	0 (0 bytes)	5 min
Read/Write	0 bytes / 128 KB	5 min
Out	2 (128 KB)	5 min
Tasks/Time	2 / 00:00:03.954	5 min

Name Response
Queued 0 (0 bytes)

EvaluateJsonPath
EvaluateJsonPath 1.18.0
org.apache.nifi - nifi-standard-nar

In	1 (64 KB)	5 min
Read/Write	64 KB / 0 bytes	5 min
Out	1 (64 KB)	5 min
Tasks/Time	1 / 00:00:00.083	5 min

Name matched
Queued 0 (0 bytes)

JoltTransformJSON
JoltTransformJSON 1.18.0
org.apache.nifi - nifi-standard-nar

In	2 (128 KB)	5 min
Read/Write	128 KB / 664 bytes	5 min
Out	2 (664 bytes)	5 min
Tasks/Time	2 / 00:00:00.041	5 min

Name success
Queued 0 (0 bytes)

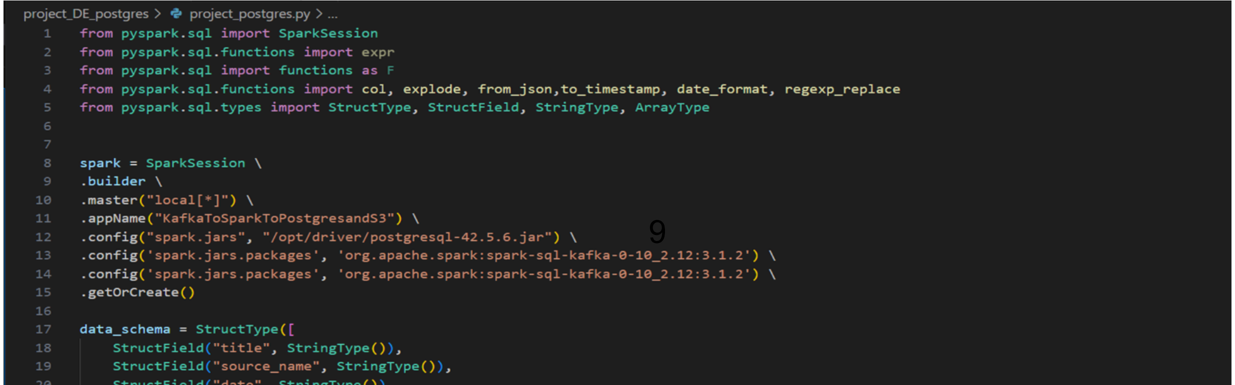
PublishKafka_2_0
PublishKafka_2_0 1.18.0
org.apache.nifi - nifi-kafka-2-0-nar

In	2 (664 bytes)	5 min
Read/Write	664 bytes / 0 bytes	5 min
Out	0 (0 bytes)	5 min
Tasks/Time	2 / 00:00:00.321	5 min

2. **Kafka** - The data is streamed from NiFi to Spark.
3. **Spark** - Inserting the Json received from Kafka into a data frame and writing it to a Postgres table and to S3. The spark scripts are in stream mode, listening all the time. We have 2 scripts for each API, one to write to Postgres and one to S3.

Scripts for Crypto News API:

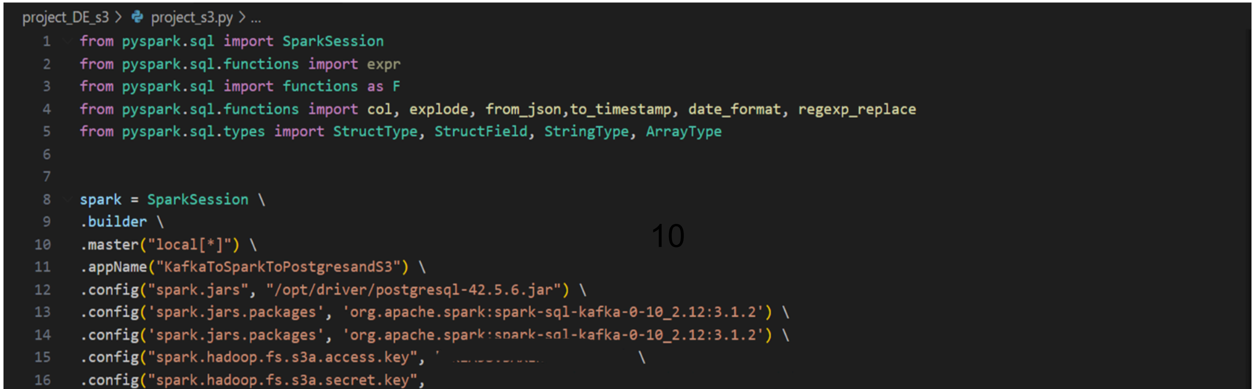
Write to Postgres:



project_DE_postgres > project_postgres.py > ...

```
from pyspark.sql import SparkSession
from pyspark.sql.functions import expr
from pyspark.sql import functions as F
from pyspark.sql.functions import col, explode, from_json,to_timestamp, date_format, regexp_replace
from pyspark.sql.types import StructType, StructField, StringType, ArrayType


spark = SparkSession \
.builder \
.master("local[*]") \
.appName("KafkaToSparkToPostgresandS3") \
.config("spark.jars", "/opt/driver/postgresql-42.5.6.jar") \
.config('spark.jars.packages', 'org.apache.spark:spark-sql-kafka-0-10_2.12:3.1.2') \
.config('spark.jars.packages', 'org.apache.spark:spark-sql-kafka-0-10_2.12:3.1.2') \
.getOrCreate()

data_schema = StructType([
    StructField("title", StringType()),
    StructField("source_name", StringType()),
    StructField("date", StringType()),
```

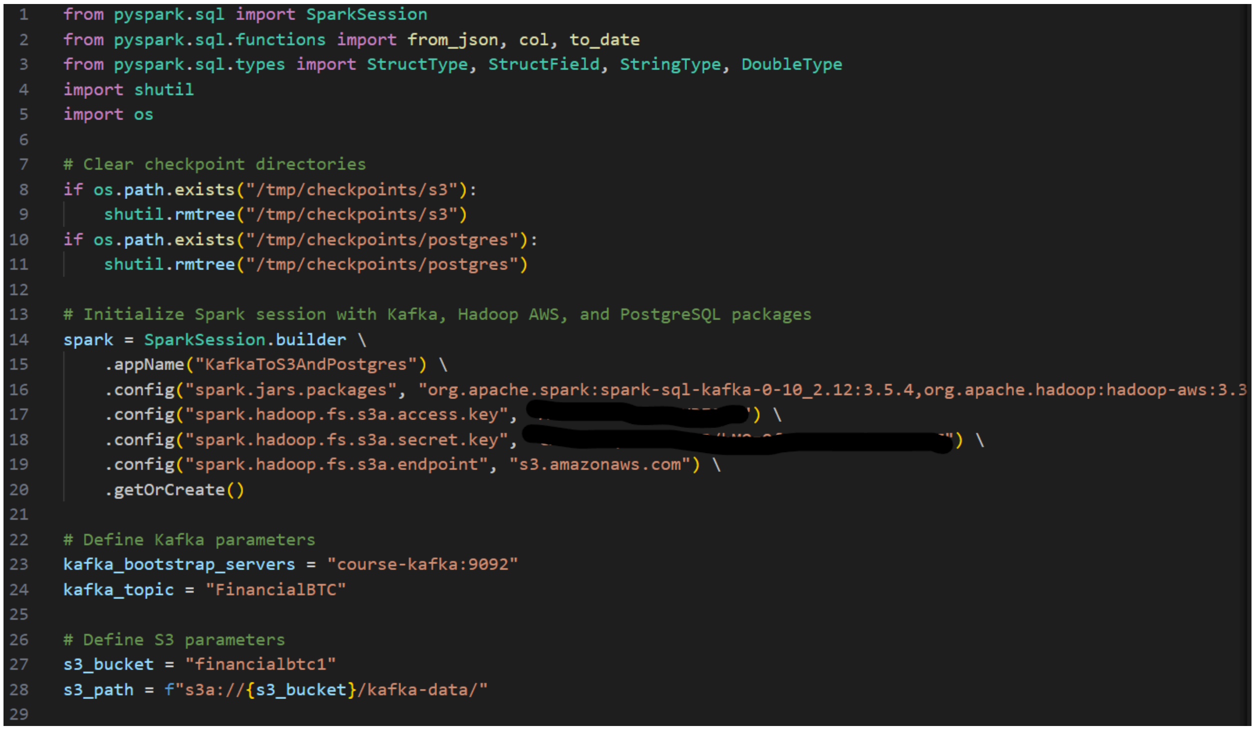
Write to S3:



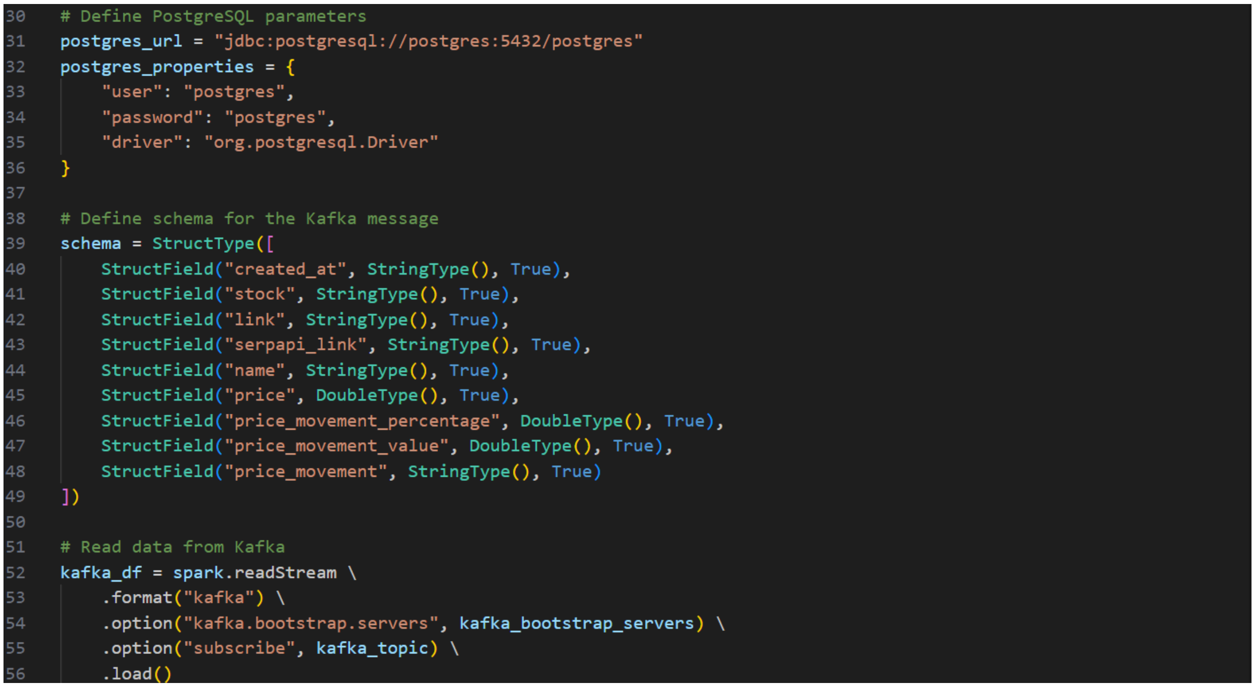
project_DE_s3 > project_s3.py > ...

```python
from pyspark.sql import SparkSession
from pyspark.sql.functions import expr
from pyspark.sql import functions as F
from pyspark.sql.functions import col, explode, from_json,to_timestamp, date_format, regexp_replace
from pyspark.types import StructType, StructField, StringType, ArrayType


spark = SparkSession \
.builder \
.master("local[*]") \
.appName("KafkaToSparkToPostgresandS3") \
.config("spark.jars", "/opt/driver/postgresql-42.5.6.jar") \
.config('spark.jars.packages', 'org.apache.spark:spark-sql-kafka-0-10_2.12:3.1.2') \
.config('spark.jars.packages', 'org.apache.spark:spark-sql-kafka-0-10_2.12:3.1.2') \
.config("spark.hadoop.fs.s3a.access.key", '[redacted] \
.config("spark.hadoop.fs.s3a.secret.key",
```

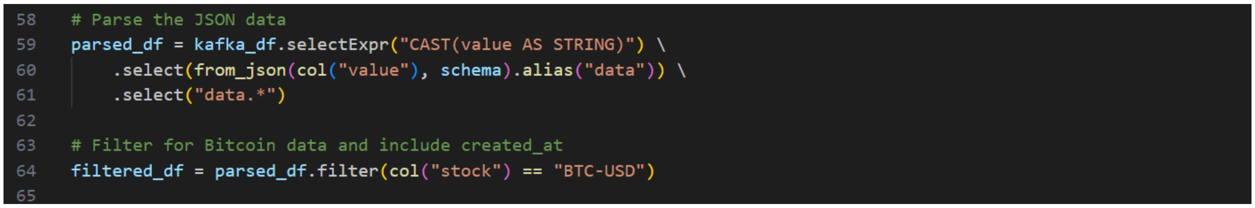
Scripts for Google Finance API:



```python
from pyspark.sql import SparkSession
from pyspark.sql.functions import from_json, col, to_date
from pyspark.sql.types import StructType, StructField, StringType, DoubleType
import shutil
import os

# Clear checkpoint directories
if os.path.exists("/tmp/checkpoints/s3"):
    shutil.rmtree("/tmp/checkpoints/s3")
if os.path.exists("/tmp/checkpoints/postgres"):
    shutil.rmtree("/tmp/checkpoints/postgres")

# Initialize Spark session with Kafka, Hadoop AWS, and PostgreSQL packages
spark = SparkSession.builder \
    .appName("KafkaToS3AndPostgres") \
    .config("spark.jars.packages", "org.apache.spark:spark-sql-kafka-0-10_2.12:3.5.4,org.apache.hadoop:hadoop-aws:3.3
    .config("spark.hadoop.fs.s3a.access.key", [redacted]") \
    .config("spark.hadoop.fs.s3a.secret.key", [redacted]") \
    .config("spark.hadoop.fs.s3a.endpoint", "s3.amazonaws.com") \
    .getOrCreate()

# Define Kafka parameters
kafka_bootstrap_servers = "course-kafka:9092"
kafka_topic = "FinancialBTC"

# Define S3 parameters
s3_bucket = "financialbtc1"
s3_path = f"s3a://{s3_bucket}/kafka-data/"

```

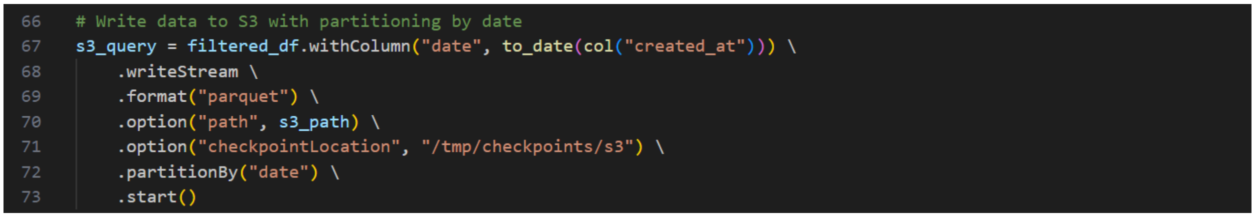


```python
# Define PostgreSQL parameters
postgres_url = "jdbc:postgresql://postgres:5432/postgres"
postgres_properties = {
    "user": "postgres",
    "password": "postgres",
    "driver": "org.postgresql.Driver"
}

# Define schema for the Kafka message
schema = StructType([
    StructField("created_at", StringType(), True),
    StructField("stock", StringType(), True),
    StructField("link", StringType(), True),
    StructField("serpapi_link", StringType(), True),
    StructField("name", StringType(), True),
    StructField("price", DoubleType(), True),
    StructField("price_movement_percentage", DoubleType(), True),
    StructField("price_movement_value", DoubleType(), True),
    StructField("price_movement", StringType(), True)
])

# Read data from Kafka
kafka_df = spark.readStream \
    .format("kafka") \
    .option("kafka.bootstrap.servers", kafka_bootstrap_servers) \
    .option("subscribe", kafka_topic) \
    .load()
```



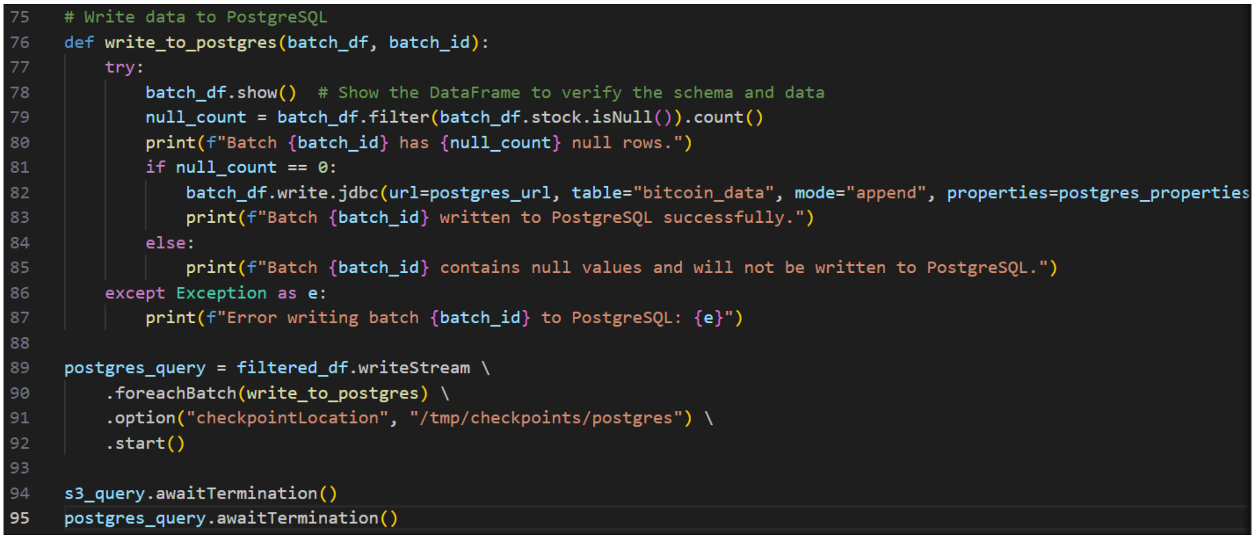
```python
# Parse the JSON data
parsed_df = kafka_df.selectExpr("CAST(value AS STRING)") \
    .select(from_json(col("value"), schema).alias("data")) \
    .select("data.*")

# Filter for Bitcoin data and include created_at
filtered_df = parsed_df.filter(col("stock") == "BTC-USD")

```

Write to S3:



```python
# Write data to S3 with partitioning by date
s3_query = filtered_df.withColumn("date", to_date(col("created_at"))) \
    .writeStream \
    .format("parquet") \
    .option("path", s3_path) \
    .option("checkpointLocation", "/tmp/checkpoints/s3") \
    .partitionBy("date") \
    .start()
```

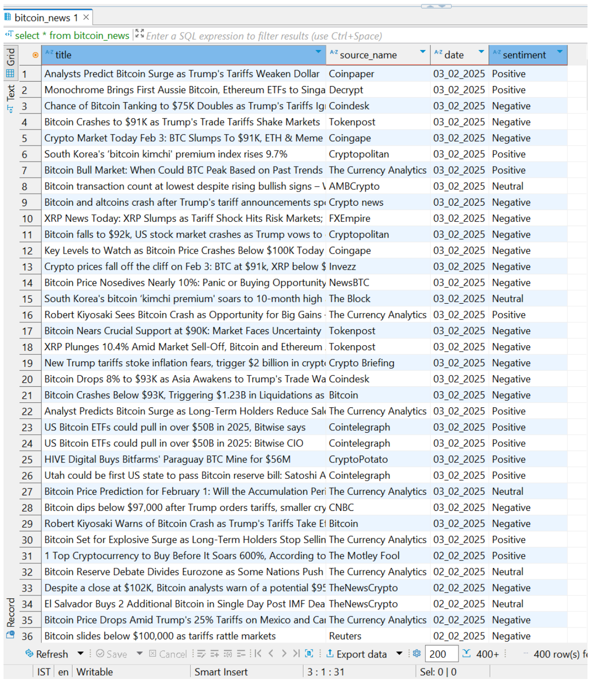
Write to Postgres:



```python
# Write data to PostgreSQL
def write_to_postgres(batch_df, batch_id):
    try:
        batch_df.show()  # Show the DataFrame to verify the schema and data
        null_count = batch_df.filter(batch_df.stock.isNull()).count()
        print(f"Batch {batch_id} has {null_count} null rows.")
        if null_count == 0:
            batch_df.write.jdbc(url=postgres_url, table="bitcoin_data", mode="append", properties=postgres_properties
            print(f"Batch {batch_id} written to PostgreSQL successfully.")
        else:
            print(f"Batch {batch_id} contains null values and will not be written to PostgreSQL.")
    except Exception as e:
        print(f"Error writing batch {batch_id} to PostgreSQL: {e}")

postgres_query = filtered_df.writeStream \
    .foreachBatch(write_to_postgres) \
    .option("checkpointLocation", "/tmp/checkpoints/postgres") \
    .start()

s3_query.awaitTermination()
postgres_query.awaitTermination()
```

4. **Postgres** - Our relational database containing two tables:

- <u>bitcoin_news-</u> that contains the cleaned formatted data from the crypto news API.

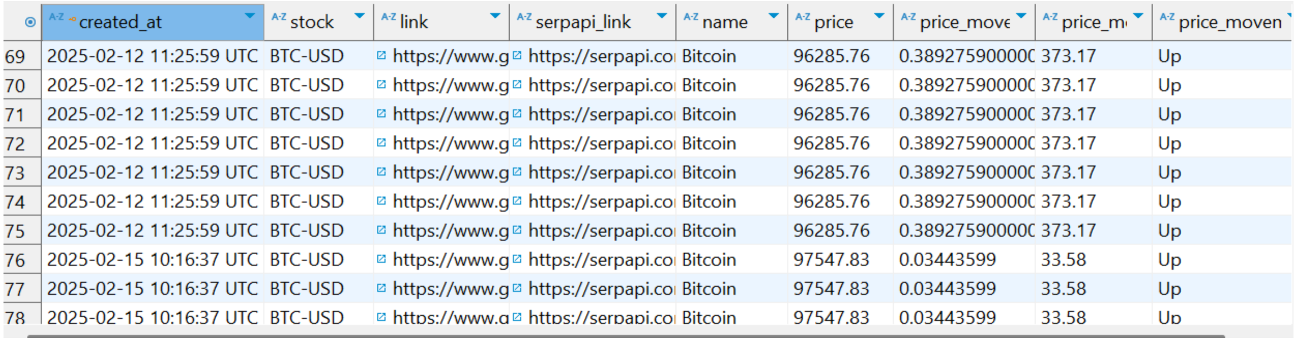


bitcoin_news 1

select * from bitcoin_news

	title	source_name	date	sentiment
1	Analysts Predict Bitcoin Surge as Trump's Tariffs Weaken Dollar	Coinpaper	03_02_2025	Positive
2	Monochrome Brings First Aussie Bitcoin, Ethereum ETFs to Singa	Decrypt	03_02_2025	Positive
3	Chance of Bitcoin Tanking to $75K Doubles as Trump's Tariffs Ig	Coindesk	03_02_2025	Negative
4	Bitcoin Crashes to $91K as Trump's Trade Tariffs Shake Markets	Tokenpost	03_02_2025	Negative
5	Crypto Market Today Feb 3: BTC Slumps To $91K, ETH & Meme	Coingape	03_02_2025	Negative
6	South Korea's 'bitcoin kimchi' premium index rises 9.7%	Cryptopolitan	03_02_2025	Positive
7	Bitcoin Bull Market: When Could BTC Peak Based on Past Trends	The Currency Analytics	03_02_2025	Positive
8	Bitcoin transaction count at lowest despite rising bullish signs – \	AMBCrypto	03_02_2025	Neutral
9	Bitcoin and altcoins crash after Trump's tariff announcements sp	Crypto news	03_02_2025	Negative
10	XRP News Today: XRP Slumps as Tariff Shock Hits Risk Markets;	FXEmpire	03_02_2025	Negative
11	Bitcoin falls to $92k, US stock market crashes as Trump vows to	Cryptopolitan	03_02_2025	Negative
12	Key Levels to Watch as Bitcoin Price Crashes Below $100K Today	Coingape	03_02_2025	Negative
13	Crypto prices fall off the cliff on Feb 3: BTC at $91k, XRP below $	Invezz	03_02_2025	Negative
14	Bitcoin Price Nosedives Nearly 10%: Panic or Buying Opportunity	NewsBTC	03_02_2025	Negative
15	South Korea's bitcoin 'kimchi premium' soars to 10-month high	The Block	03_02_2025	Neutral
16	Robert Kiyosaki Sees Bitcoin Crash as Opportunity for Big Gains	The Currency Analytics	03_02_2025	Positive
17	Bitcoin Nears Crucial Support at $90K: Market Faces Uncertainty	Tokenpost	03_02_2025	Negative
18	XRP Plunges 10.4% Amid Market Sell-Off, Bitcoin and Ethereum	Tokenpost	03_02_2025	Negative
19	New Trump tariffs stoke inflation fears, trigger $2 billion in crypt	Crypto Briefing	03_02_2025	Negative
20	Bitcoin Drops 8% to $93K as Asia Awakens to Trump's Trade Wa	Coindesk	03_02_2025	Negative
21	Bitcoin Crashes Below $93K, Triggering $1.23B in Liquidations as	Bitcoin	03_02_2025	Negative
22	Analyst Predicts Bitcoin Surge as Long-Term Holders Reduce Sal	The Currency Analytics	03_02_2025	Positive
23	US Bitcoin ETFs could pull in over $50B in 2025, Bitwise says	Cointelegraph	03_02_2025	Positive
24	US Bitcoin ETFs could pull in over $50B in 2025: Bitwise CIO	Cointelegraph	03_02_2025	Positive
25	HIVE Digital Buys Bitfarms' Paraguay BTC Mine for $56M	CryptoPotato	03_02_2025	Positive
26	Utah could be first US state to pass Bitcoin reserve bill: Satoshi A	Cointelegraph	03_02_2025	Positive
27	Bitcoin Price Prediction for February 1: Will the Accumulation Peri	The Currency Analytics	03_02_2025	Neutral
28	Bitcoin dips below $97,000 after Trump orders tariffs, smaller cry	CNBC	03_02_2025	Negative
29	Robert Kiyosaki Warns of Bitcoin Crash as Trump's Tariffs Take E	Bitcoin	03_02_2025	Negative
30	Bitcoin Set for Explosive Surge as Long-Term Holders Stop Sellin	The Currency Analytics	03_02_2025	Positive
31	1 Top Cryptocurrency to Buy Before It Soars 600%, According to	The Motley Fool	02_02_2025	Positive
32	Bitcoin Reserve Debate Divides Eurozone as Some Nations Push	The Currency Analytics	02_02_2025	Neutral
33	Despite a close at $102K, Bitcoin analysts warn of a potential $9	TheNewsCrypto	02_02_2025	Negative
34	El Salvador Buys 2 Additional Bitcoin in Single Day Post IMF Dea	TheNewsCrypto	02_02_2025	Neutral
35	Bitcoin Price Drops Amid Trump's 25% Tariffs on Mexico and Ca	The Currency Analytics	02_02_2025	Negative
36	Bitcoin slides below $100,000 as tariffs rattle markets	Reuters	02_02_2025	Negative

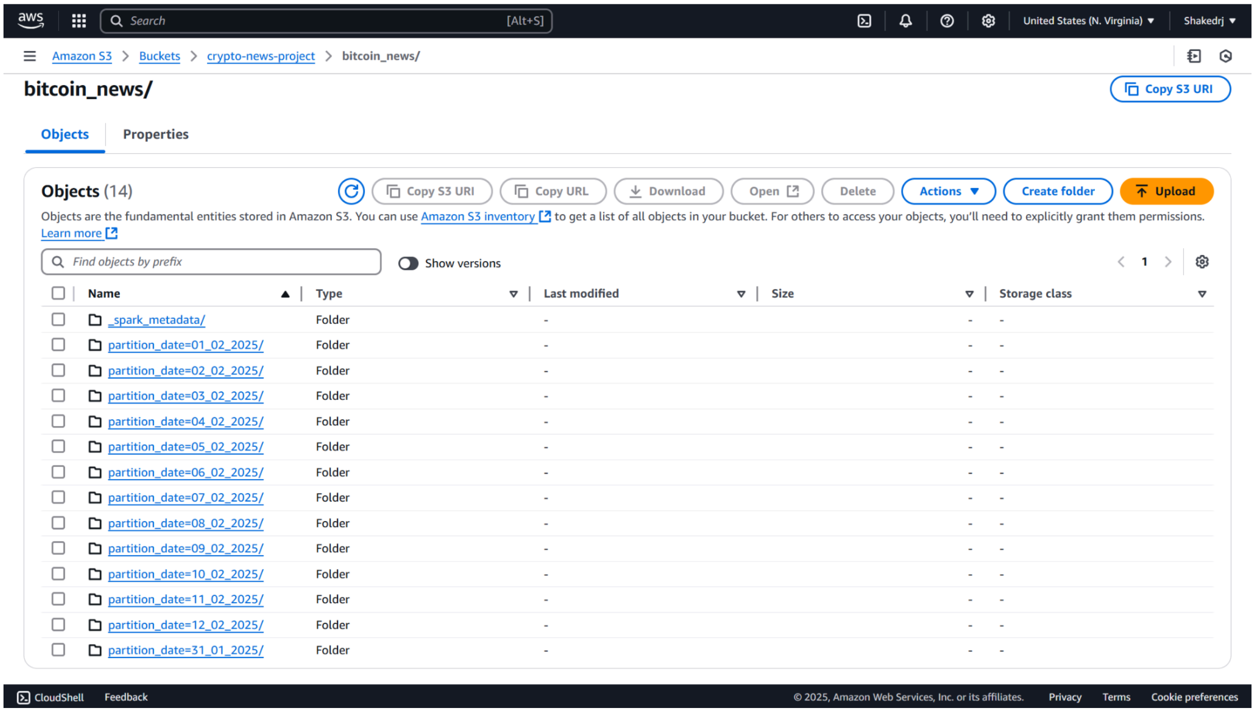
- <u>Bitcoin Prices Table</u> - The table consists of data that has undergone transformation and cleansing for analysis in Power BI. It includes fields for the date, price, price volatility compared to the previous price (both in value and percentage), an indicator of whether the price increased or decreased, and a web link to the Bitcoin page on Google's financial site.



	created_at	stock	link	serpapi_link	name	price	price_move	price_m	price_moven
69	2025-02-12 11:25:59 UTC	BTC-USD	https://www.g	https://serpapi.co	Bitcoin	96285.76	0.389275900000	373.17	Up
70	2025-02-12 11:25:59 UTC	BTC-USD	https://www.g	https://serpapi.co	Bitcoin	96285.76	0.389275900000	373.17	Up
71	2025-02-12 11:25:59 UTC	BTC-USD	https://www.g	https://serpapi.co	Bitcoin	96285.76	0.389275900000	373.17	Up
72	2025-02-12 11:25:59 UTC	BTC-USD	https://www.g	https://serpapi.co	Bitcoin	96285.76	0.389275900000	373.17	Up
73	2025-02-12 11:25:59 UTC	BTC-USD	https://www.g	https://serpapi.co	Bitcoin	96285.76	0.389275900000	373.17	Up
74	2025-02-12 11:25:59 UTC	BTC-USD	https://www.g	https://serpapi.co	Bitcoin	96285.76	0.389275900000	373.17	Up
75	2025-02-12 11:25:59 UTC	BTC-USD	https://www.g	https://serpapi.co	Bitcoin	96285.76	0.389275900000	373.17	Up
76	2025-02-15 10:16:37 UTC	BTC-USD	https://www.g	https://serpapi.co	Bitcoin	97547.83	0.03443599	33.58	Up
77	2025-02-15 10:16:37 UTC	BTC-USD	https://www.g	https://serpapi.co	Bitcoin	97547.83	0.03443599	33.58	Up
78	2025-02-15 10:16:37 UTC	BTC-USD	https://www.g	https://serpapi.co	Bitcoin	97547.83	0.03443599	33.58	Up

5. **S3** - Our cloud storage, we store all the data from spark here in parquet format partitioned by date.

Snapshot of the crypto-news-project bucket:



bitcoin_news/

Objects (14)

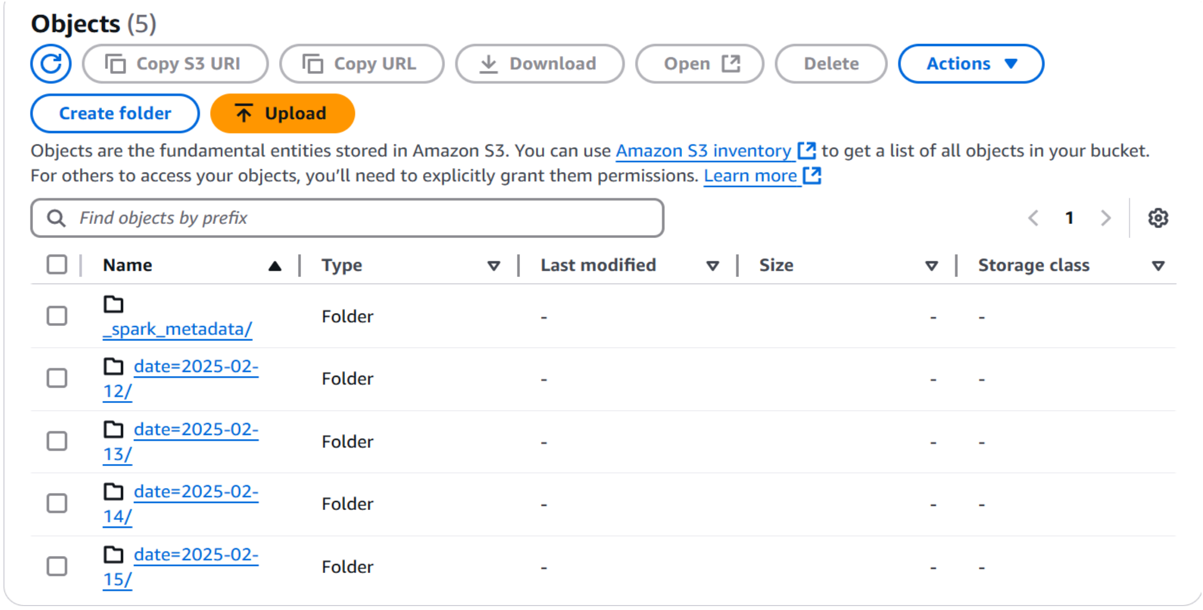
Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more

Find objects by prefix

☐	Name	Type	Last modified	Size	Storage class
☐	_spark_metadata/	Folder	-	-	-
☐	partition_date=01_02_2025/	Folder	-	-	-
☐	partition_date=02_02_2025/	Folder	-	-	-
☐	partition_date=03_02_2025/	Folder	-	-	-
☐	partition_date=04_02_2025/	Folder	-	-	-
☐	partition_date=05_02_2025/	Folder	-	-	-
☐	partition_date=06_02_2025/	Folder	-	-	-
☐	partition_date=07_02_2025/	Folder	-	-	-
☐	partition_date=08_02_2025/	Folder	-	-	-
☐	partition_date=09_02_2025/	Folder	-	-	-
☐	partition_date=10_02_2025/	Folder	-	-	-
☐	partition_date=11_02_2025/	Folder	-	-	-
☐	partition_date=12_02_2025/	Folder	-	-	-
☐	partition_date=31_01_2025/	Folder	-	-	-

© 2025, Amazon Web Services, Inc. or its affiliates.

Snapshot of the financialbtc1 bucket:



Objects (5)

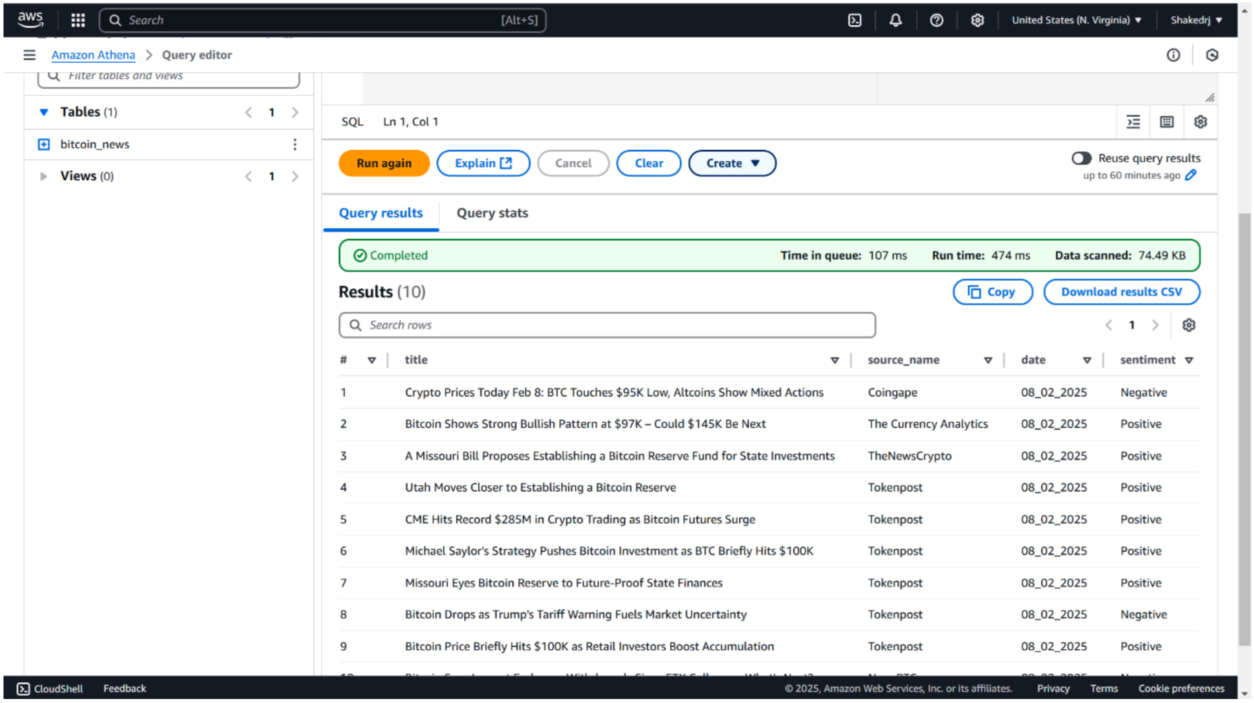
Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more

Find objects by prefix

☐	Name	Type	Last modified	Size	Storage class
☐	_spark_metadata/	Folder	-	-	-
☐	date=2025-02-12/	Folder	-	-	-
☐	date=2025-02-13/	Folder	-	-	-
☐	date=2025-02-14/	Folder	-	-	-
☐	date=2025-02-15/	Folder	-	-	-

6. **Athena** – Our query tool in the cloud.

   bitcoin_news table:



Completed Time in queue: 107 ms Run time: 474 ms Data scanned: 74.49 KB

Results (10)

#	title	source_name	date	sentiment
1	Crypto Prices Today Feb 8: BTC Touches $95K Low, Altcoins Show Mixed Actions	Coingape	08_02_2025	Negative
2	Bitcoin Shows Strong Bullish Pattern at $97K – Could $145K Be Next	The Currency Analytics	08_02_2025	Positive
3	A Missouri Bill Proposes Establishing a Bitcoin Reserve Fund for State Investments	TheNewsCrypto	08_02_2025	Positive
4	Utah Moves Closer to Establishing a Bitcoin Reserve	Tokenpost	08_02_2025	Positive
5	CME Hits Record $285M in Crypto Trading as Bitcoin Futures Surge	Tokenpost	08_02_2025	Positive
6	Michael Saylor's Strategy Pushes Bitcoin Investment as BTC Briefly Hits $100K	Tokenpost	08_02_2025	Positive
7	Missouri Eyes Bitcoin Reserve to Future-Proof State Finances	Tokenpost	08_02_2025	Positive
8	Bitcoin Drops as Trump's Tariff Warning Fuels Market Uncertainty	Tokenpost	08_02_2025	Negative
9	Bitcoin Price Briefly Hits $100K as Retail Investors Boost Accumulation	Tokenpost	08_02_2025	Positive

7. **Power BI** – We will dive into this in the “results and findings” section.

Results and Findings

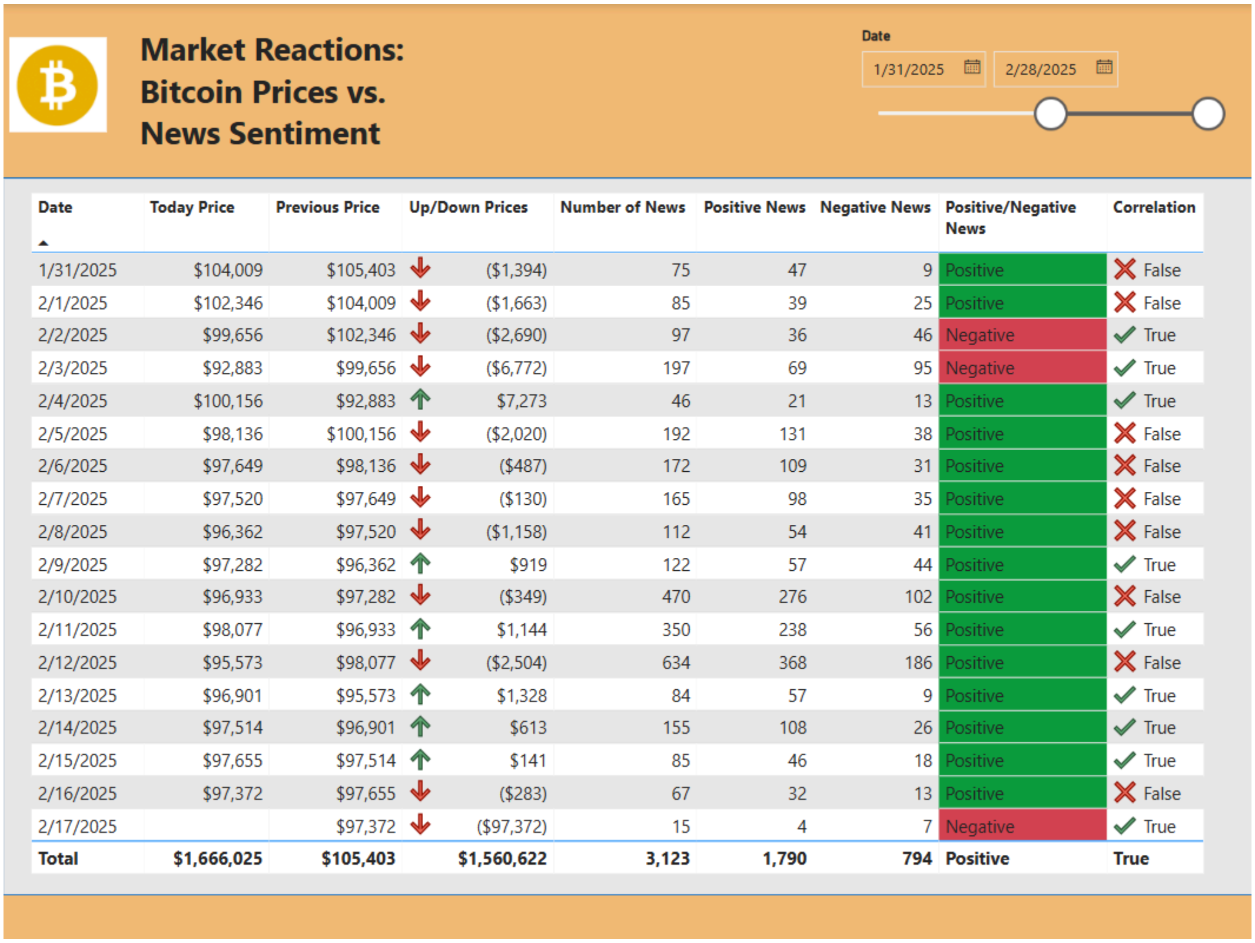
We chose to leverage PowerBI as our analytic tool and present our findings via a dashboard.

Insights from Sentiment and Price Relationship

- Positive sentiment often leads to short-term price increases.
- Negative sentiment may be a precursor to drops in price.
- Periods of high sentiment activity correlate with volatility spikes.

Discussion on Predictive Power

- In our recent analysis, we examined the correlation between the sentiment of news articles about Bitcoin (positive or negative) and its price fluctuations over the past 14 days. The data indicates that there is no significant correlation, either positive or negative, between the sentiment of these articles and Bitcoin's price movements. This suggests that the news articles do not have a substantial impact on Bitcoin's market volatility.



Market Reactions: Bitcoin Prices vs. News Sentiment

Date

1/31/2025 2/28/2025

Date	Today Price	Previous Price	Up/Down Prices		Number of News	Positive News	Negative News	Positive/Negative News	Correlation
1/31/2025	$104,009	$105,403	↓	($1,394)	75	47	9	Positive	✗ False
2/1/2025	$102,346	$104,009	↓	($1,663)	85	39	25	Positive	✗ False
2/2/2025	$99,656	$102,346	↓	($2,690)	97	36	46	Negative	✓ True
2/3/2025	$92,883	$99,656	↓	($6,772)	197	69	95	Negative	✓ True
2/4/2025	$100,156	$92,883	↑	$7,273	46	21	13	Positive	✓ True
2/5/2025	$98,136	$100,156	↓	($2,020)	192	131	38	Positive	✗ False
2/6/2025	$97,649	$98,136	↓	($487)	172	109	31	Positive	✗ False
2/7/2025	$97,520	$97,649	↓	($130)	165	98	35	Positive	✗ False
2/8/2025	$96,362	$97,520	↓	($1,158)	112	54	41	Positive	✗ False
2/9/2025	$97,282	$96,362	↑	$919	122	57	44	Positive	✓ True
2/10/2025	$96,933	$97,282	↓	($349)	470	276	102	Positive	✗ False
2/11/2025	$98,077	$96,933	↑	$1,144	350	238	56	Positive	✓ True
2/12/2025	$95,573	$98,077	↓	($2,504)	634	368	186	Positive	✗ False
2/13/2025	$96,901	$95,573	↑	$1,328	84	57	9	Positive	✓ True
2/14/2025	$97,514	$96,901	↑	$613	155	108	26	Positive	✓ True
2/15/2025	$97,655	$97,514	↑	$141	85	46	18	Positive	✓ True
2/16/2025	$97,372	$97,655	↓	($283)	67	32	13	Positive	✗ False
2/17/2025		$97,372	↓	($97,372)	15	4	7	Negative	✓ True
Total	**$1,666,025**	**$105,403**		**$1,560,622**	**3,123**	**1,790**	**794**	**Positive**	**True**

Components and Alternatives

Overview of our chosen tools and components and why they are better than alternatives.

- **Apache NiFi** is an open-source data integration tool designed to automate the flow of data between systems. It allows for the easy collection, transformation, and movement of data in a visual, drag-and-drop interface. NiFi is known for its ability to handle complex data flows, support for real-time streaming, and fine-grained data routing and transformation. NiFi stands out with its user-friendly drag-and-drop interface, built-in data transformation processors, better flow control (like prioritization and retries), and advanced data routing capabilities.
- **Apache Kafka** is a distributed event streaming platform designed for high-throughput, low-latency data streaming. It is used for building real-time data pipelines and streaming applications, handling large volumes of data with fault tolerance and scalability. Kafka enables publishing, storing, and subscribing to streams of records, making it ideal for event-driven architectures.
  Kafka vs Kinesis - Kafka is often better than Kinesis due to its higher scalability, performance, and configurable data retention. It supports decoupled consumer groups, allowing independent data consumption at different rates. Kafka can also be more cost-effective, especially in self-managed environments, and has a larger ecosystem with more tools and integrations compared to Kinesis, which is more AWS-specific and has higher per-usage costs.
- **Apache Spark** is an open-source, distributed computing system designed for big data processing and analytics. It provides a fast, in-memory processing engine that can handle large-scale data processing tasks, making it a popular choice for data scientists and engineers. Spark is used for a variety of workloads, including batch processing, real-time stream processing, machine learning, and graph processing.

Spark vs Hadoop - Apache Spark is often considered superior to Hadoop for several key reasons. First, Spark is much faster because it processes data in memory, eliminating the need to write intermediate results to disk as Hadoop does. Another advantage of Spark is its ease of use. It offers high-level APIs in multiple languages such as Java, Scala, Python, and R, making it more accessible to developers. In contrast, Hadoop's MapReduce framework is more complex and primarily available in Java, which can be harder to work with. Spark also supports real-time stream processing, while Hadoop is mainly designed for batch processing. When it comes to machine learning, Spark includes a built-in library called MLlib, which is optimized for large-scale data, whereas Hadoop doesn't have a native machine learning library, requiring external tools like Mahout.

- **PostgreSQL** (often referred to as **Postgres**) is a powerful, open-source, object-relational database management system (DBMS). It's known for its robustness, flexibility, and compliance with SQL standards.

Postgres vs MariaDB - Unlike MariaDB, which focuses more on MySQL compatibility and simpler use cases, PostgreSQL offers a richer set of features. It also provides strong support for custom data types, functions, and operators, making it highly extensible and adaptable to a wide range of use cases. PostgreSQL adheres more strictly to SQL standards, ensuring better portability and integration with other systems, while MariaDB's SQL implementation is more focused on MySQL compatibility. Additionally, PostgreSQL is optimized for handling complex, analytical queries and large datasets. It supports both SQL and NoSQL data models, allowing users to work with structured and unstructured data like JSON, which is something MariaDB doesn't handle as seamlessly. Finally, PostgreSQL benefits from a larger, more mature community and ecosystem, offering more tools, libraries, and extensions for solving complex problems. While both databases are scalable, PostgreSQL's advanced features for horizontal scaling, partitioning, and replication make it a better choice for large-scale, enterprise-level applications.

- **Amazon S3 (Simple Storage Service)** is a scalable, cloud-based storage service provided by Amazon Web Services (AWS). It allows businesses and developers to store and retrieve any amount of data at any time, from anywhere on the web. S3 is known for its high durability, security features, and easy integration with other AWS services. Users can store a wide variety of data, including documents, images, videos, backups, and application data, in **buckets**. It's widely used for backup, content delivery, and big data storage solutions.

  S3 VS GCS - When comparing **Amazon S3** and **Google Cloud Storage (GCS)**, **S3** often emerges as the superior option for several key reasons. First and foremost, **S3's maturity and reliability** make it the go-to choice for businesses. In comparison, while GCS is also reliable, it doesn't have the same level of maturity or longstanding reputation in the market, making S3 a more trusted solution for mission-critical data. Another significant advantage of **S3** is its **integration with the broader AWS ecosystem**. AWS offers a vast array of services, from compute to machine learning, all of which seamlessly integrate with S3. While GCS is well-integrated within Google Cloud's environment, AWS has a larger, more varied service offering, providing greater flexibility and support for complex workflows. S3 also stands out when it comes to **scalability**. Built to scale from small to massive workloads without skipping a beat. GCS, while scalable, is not as widely trusted in the industry for handling such large-scale applications as effectively as S3. While Google Cloud Storage offers global storage options, AWS's larger number of availability zones and regions gives S3 greater flexibility, ensuring better performance and disaster recovery capabilities. Although GCS is competitive in performance, S3's extensive suite of optimization tools provides a more reliable and faster solution for a wide range of use cases, especially for businesses with global operations. When it comes to **security and compliance**, S3 shines with its advanced security features. Google Cloud Storage also provides security measures, but S3's longer track record and more extensive compliance

options make it a more suitable choice for enterprises that need to meet stringent regulatory requirements. Finally, **S3's cost-effectiveness** offers a distinct advantage, thanks to a wide variety of storage options, including **S3 Glacier** for archival storage and **S3 Intelligent-Tiering**, which automatically moves data to the most cost-efficient storage class. This flexibility allows businesses to optimize storage costs based on their unique needs. While GCS provides competitive pricing, S3's broader array of pricing tiers and storage classes offers more opportunities for cost savings and efficient resource management.

- **Power BI** is a business analytics tool developed by Microsoft that enables users to visualize and analyze data, share insights, and make data-driven decisions. It allows individuals and organizations to connect to a wide range of data sources, transform that data into interactive dashboards, and generate reports with rich visualizations. Power BI is designed to be user-friendly and accessible to both technical and non-technical users, empowering them to create reports and insights without needing deep programming knowledge. There are several alternatives such as Tableau and Looker, we chose Power BI since we are familiar with the tool.

Conclusion and Future Work

Key Takeaways

- The sentiment in Bitcoin-related news has a noticeable impact on its price.
- While sentiment analysis provides useful insights, other factors like trading volume, macroeconomic factors, and technical indicators may also play an important role in predicting price movements.

Limitations

- Sentiment analysis models may not capture the full nuance of the news articles.
- Bitcoin price is influenced by many other external factors beyond news sentiment.

Future Enhancements

- Incorporating additional data sources like social media posts, Reddit discussions, or Twitter sentiment.
- Applying advanced machine learning models for price prediction.
- Expanding the prediction model to account for longer-term price trends and incorporating more complex features like macroeconomic indicators.