



Politechnika Warszawska

Crypto Options vs. Rates

Jan Skwarek, Sebastian
Pergała, Aleksander
Malinowski

Analiza Korelacji: Rynki Predykcyjne (Polymarket)
vs. Ceny Spot (Binance)

14.01.2026



Analiza: Sentyment vs. Rzeczywistość

Cel: Badanie korelacji między zakładami (Polymarket) a kursem kryptowalut (Binance)

Źródła: Dane z rynków predykcyjnych oraz rzeczywiste ceny spot.

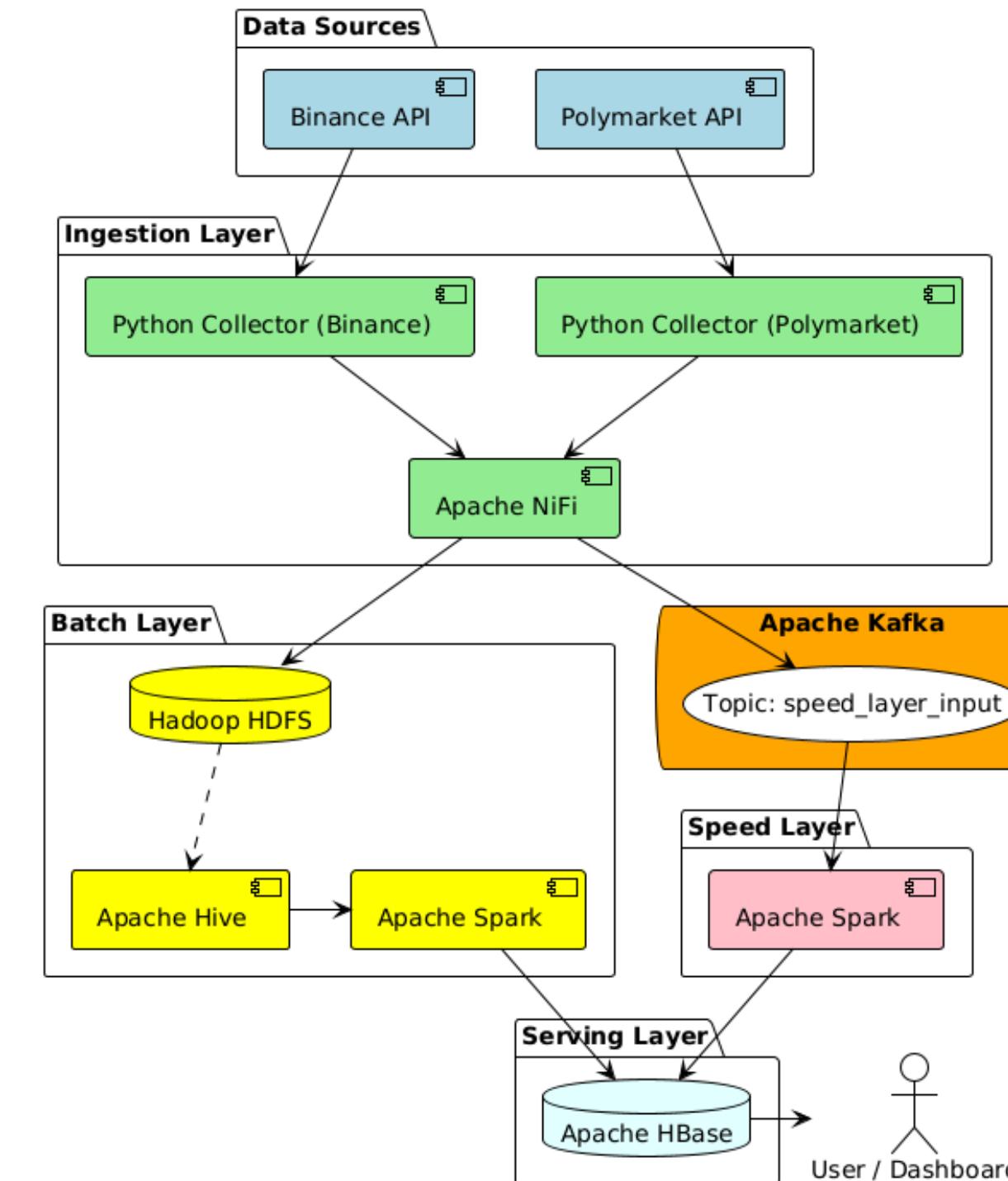
Podejście: Wykorzystanie "mądrości tłumu" do zrozumienia ruchów cenowych.

Technologia: Skalowalne przetwarzanie w architekturze Lambda.





Architektura systemu





Środowisko i automatyzacja

```
vagrant@node1: ~/Crypto-Op ~ + - X
To verify, run: hbase shell
  > describe 'market_analytics'
  > describe 'market_live'
=====
=====
Initialization Complete!
=====

Available Commands:
- Start ingestion: bash console_scripts/start_ingestion.sh
- Stop ingestion: bash console_scripts/stop_ingestion.sh
- Batch cron status: bash console_scripts/setup_batch_cron.sh status
- View batch logs: bash console_scripts/setup_batch_cron.sh logs
- Run batch manually: bash batch_layer/run_batch_analytics.sh
=====

vagrant@node1:~/Crypto-Options-vs-Rates$ console_scripts/start_ingestion.sh
=====
Starting Ingestion & Speed Layer
=====
[CHECK] Verifying Infrastructure status...
-> HDFS is running and reachable.
[INGEST] Starting Binance Collector...
-> Binance Collector running (PID: 15175)
[INGEST] Starting Polymarket WebSocket...
-> Polymarket CLOB Collector running (PID: 15177)
[SPEED] Starting Spark Speed Layer...
```



Crypto Options vs. Rates

Źródła danych

Polymarket / [How it works](#) [Log In](#) [Sign Up](#) ≡

[Trending](#) [Breaking](#) [New](#) | [Politics](#) [Sports](#) [Crypto](#) [Finance](#) [Geopolitics](#) [Earnings](#) [Tech](#) [Culture](#) [World](#) [Economy](#) [Climate & Science](#) [Elections](#) [Me](#)

[All](#) [Trump](#) [Iran](#) [NFL Playoffs](#) [Greenland](#) [Cellular Outage](#) [Tariffs](#) [Portugal Election](#) [Fed](#) [Derivatives](#) [Venezuela](#) [Ukraine](#) [Oscars](#) [Epst](#) > Clear filters

Sort by: [24hr Volume](#) [Frequency](#): All [Status](#): Active Hide sports? Hide crypto? Hide earnings? [Clear filters](#)

Event	Date	Probability	Yes	No	Vol.
Khamenei out as Supreme Leader of Iran...?		22%	Yes	No	\$27m Vol.
Will the Iranian regime fall before 2027?		48%	Yes	No	\$2m Vol.
US strikes Iran by...?	January 18	57%	Yes	No	
US strikes Iran by...?	January 23	60%	Yes	No	\$37m Vol.
Supreme Court rules in favor of Trump's tariffs?		31%	Yes	No	\$3m Vol.
Will Trump acquire Greenland before 2027?		18%	Yes	No	\$8m Vol.
Israel strikes Iran by January 31, 2026?		47%	Yes	No	\$9m Vol.
Portugal Presidential Election	António José Seguro... João Cotrim Figueiredo...	62% 21%	Yes Yes	No No	\$102m Vol.
Jerome Powell federally charged by June 30?		12%	Yes	No	\$96m Vol.
Jerome Powell out as Fed Chair by...?	March 31 May 14	5% 9%	Yes Yes	No No	\$372k Vol.
Who will be the first to leave the Trump Cabinet?	Pam Bondi Kristi Noem	30% 14%	Yes Yes	No No	\$879k Vol.
Who will Trump nominate as Fed Chair?	Kevin Warsh Kevin Hassett	41% 35%	Yes Yes	No No	\$185m Vol.
US next strikes Iran on...?	January 14 January 15	19% 16%	Yes Yes	No No	\$3m Vol.

BINANCE [Buy Crypto](#) [Markets](#) [Trade](#) [Futures](#) [Earn](#) [Square](#) [More](#)

304,541,402 USERS TRUST US

The World's Leading Cryptocurrency Exchange

No.1 Customer Assets **No.1** Trading Volume



Źródło danych - Binance

- **Backfill: Binance API → JSON → Parquet (HDFS)** - zaciągamy automatycznie 1-minutowe świece (klines) z ostatnich 24h, żeby mieć dane do analityki już na starcie (zapobiegamy 'cold start')
- **Stream: WebSocket → Bufor → Parquet (Live)** - połączenie przez WebSocket dla bieżących aktualizacji świec co minutę.



Standaryzacja danych

- Rejestrujemy pełne dane OHLCV
- Składujemy pliki Parquet, do których dane są zapisywane z partycjonowaniem po dacie i symbolu waluty, żeby optymalizować odczyt.

Niezawodność

- Mechanizm self-healing dla połączeń
- Deduplikacja danych przed zapisem do strefy Raw w HDFS



Źródło danych - Polymarket

Charakterystyka danych

Typ: Zdecentralizowany rynek predykcyjny.

Zakres: Krótkoterminowe rynki 15-minutowe (opcje binarne góra/dół).

Aktywa: ETH, SOL, XRP oraz BNB.

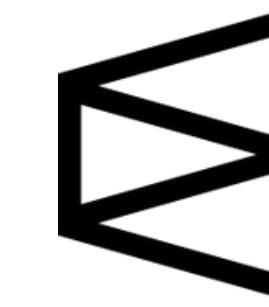
Wiarygodność: Wyniki rozliczane przez wyrocznie Chainlink.

Metoda akwizycji

Gamma API (REST): Dynamiczne odkrywanie identyfikatorów aktywnych rynków.

CLOB API (WebSocket): Odczyt strumienia zdarzeń w czasie rzeczywistym.

Przetwarzane zdarzenia: Arkusz zleceń (book), zmiany cen oraz ostatnie transakcje.



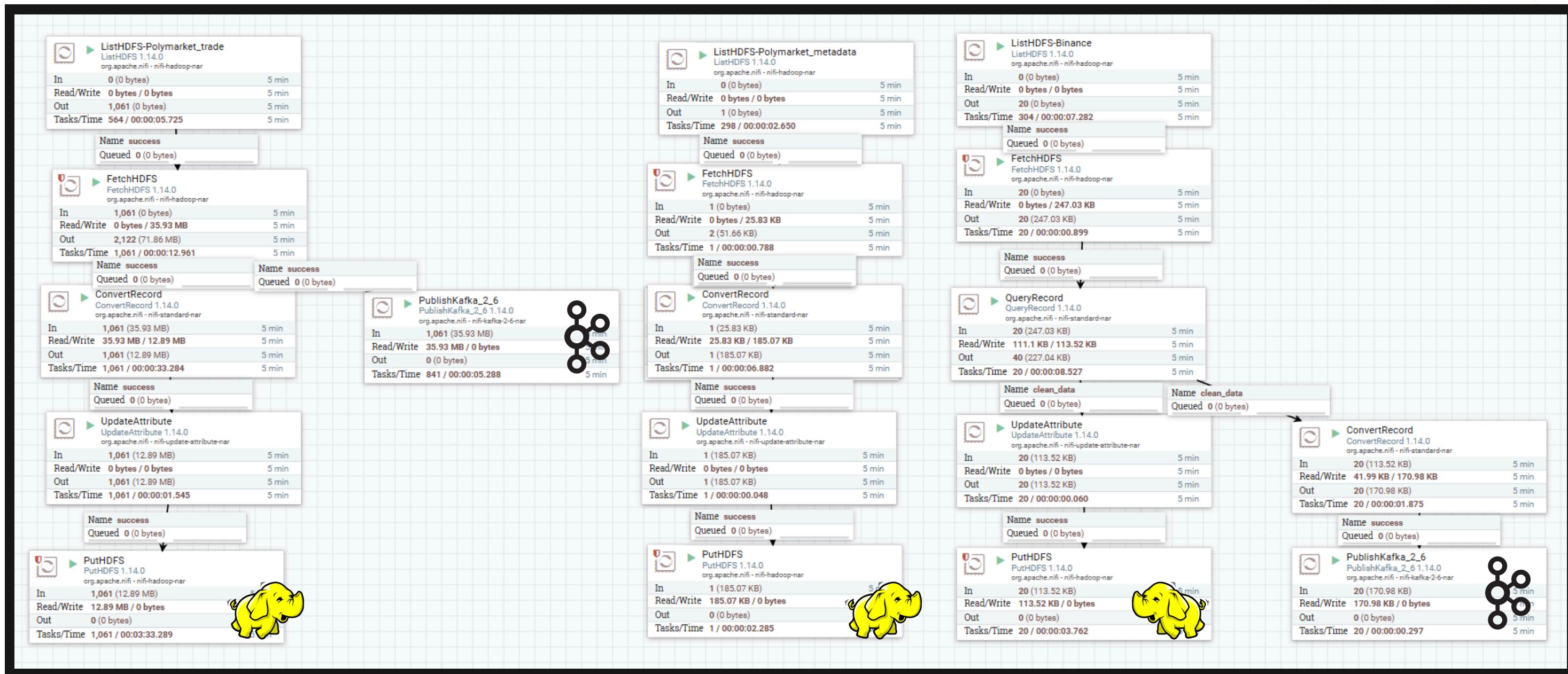
Polymarket

```
Response
{
  "market": "0x5f65177b394277fd294cd75650044e32ba009a95022d88a0c1d565897d72f8f1",
  "price_changes": [
    {
      "asset_id": "7132104567925221259462638553270691275033272857194253228963",
      "price": "0.5",
      "size": "200",
      "side": "BUY",
      "hash": "56621a121a47ed9333273e21c83b660cff37ae50",
      "best_bid": "0.5",
      "best_ask": "1"
    },
    {
      "asset_id": "5211431950124591551605510604688420996992612748282795467444",
      "price": "0.5",
      "size": "200",
      "side": "SELL",
      "hash": "1895759e4df7a796bf4f1c5a5950b748306923e2",
      "best_bid": "0",
      "best_ask": "0.5"
    }
  ],
  "timestamp": "1757908892351",
  "event_type": "price_change"
}
```



Crypto Options vs. Rates

Ingestion Layer - NiFi



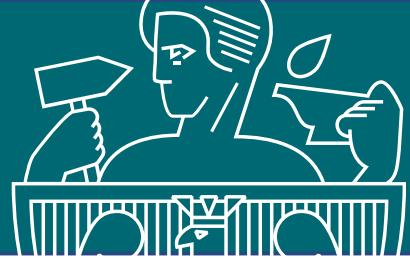


Batch Layer - PySpark

Przetwarzanie historyczne i analityka

- **Technologie:** Apache Spark (PySpark) oraz Apache Hive.
- **Harmonogram:** Cykliczne uruchamianie co 6 godzin.
- **Główne operacje:**
 - Agregacja danych do spójnych okien 15-minutowych.
 - Łączenie (Join) danych giełdowych z przewidywaniemi rynku.
 - Obliczanie korelacji i klasyfikacja trafności sygnałów.
- **Wynik:** Pre-komputowane widoki zapisywane w HBase.





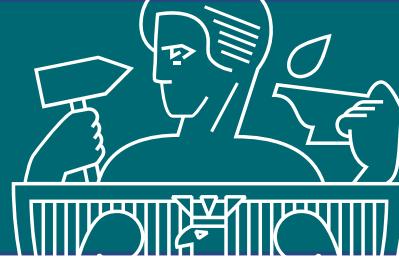
Speed Layer - Kafka



■ ■ binance

- last_trade_price
 - price_change
 - book

■ ■ polymarket_trade



Speed Layer - Spark

Binance

Column	Type
window	struct
symbol	string
current_price	double
current_sentiment	double
avg_price	double
min_price	double
max_price	double
volatility	double
total_usdt_volume	double
avg_sentiment	double
ticks	long

Polymarket

Column	Type
window	struct
market_id	string
current_prob	double
current_spread	double
current_imbalance	double
avg_prob	double
min_prob	double
max_prob	double
avg_imbalance	double
avg_spread	double
num_updates	long



Column	Type
window	struct
market_id	string
total_shares_traded	double
num_trades	long



Serving Layer HBase - integracja architektury Lambda

Schemat Tabeli

- **market_analytics** (Warstwa Batch): Przechowuje w pełni przetworzone, zagregowane okna 15-minutowe
- **market_live** (Warstwa Speed): Przechowuje sub-sekundowe aktualizacje strumieniowe ze Spark Streaming.
- **Strategia RowKey**: Zastosowaliśmy odwrócone znaczniki czasu (Reverse Timestamps).
- **Widok Scalony** (Merge View): Mechanizm łączący historyczną dokładność (Batch) z bieżącą aktualnością (Speed) w spójną odpowiedź API.

cel →

Wymuszenie fizycznego składowania najnowszych danych na początku tabeli, co umożliwia natychmiastowe pobieranie ich przez analitykę i/lub dashboardy.



Analityka i Walidacja Rozwiązania

Logika etykietowania

- Korelacja Sentymentu Rynku (kursy zakładów) z Rzeczywistością (kurs giełdowy) w oknach 15-minutowych.
- Klasifikacja wyników na: CORRECT_BULL (trafnny wzrost), FAILED_BEAR (nietraffiony spadek), UNCERTAIN (niepewny).

Manualna Walidacja

- Przeprowadzono ręczną weryfikację konkretnych okien czasowych (np. 12:00-12:15).
- Potwierdzono, że okna z wysokim pstwem (>60%) zostały poprawnie oflagowane przez Spark Batch w momentach, gdy cena rynkowa podążyła za predykcją.

Automatyzacja i Testowanie

- Testy E2E: Zweryfikowano pełny przepływ danych: WebSocket Binance → HDFS -> Spark -> HBase.
- Integralność Danych: Osiągnięto pełną spójność rekordów między źródłem (Hive) a celem (HBase) podczas testów obciążeniowych.
- Unit testing: Każdy z modułów kodu został pokryty testami jednostkowymi, odizolowanymi od środowiska, na którym działa program.

Dziękujemy
za uwagę

