

ZESPÓŁ: PROBLEM, PLAN, PROBLEMS

I'VE GOT SOME NEWS!

Projekt zrealizowany w ramach przedmiotu SDwsBD

PLAN PREZENTACJI

- Temat projektu
- Wykorzystywane dane
- Schemat rozwiązania
- Pozyskiwanie i przetwarzanie danych
- Składowanie danych (batch layer)
- Przetwarzanie wsadowe i składowanie widoków (serving layer)
- Podsumowanie i wnioski



TEMATYKA

Artykuły, Wydawnictwa, Twitter...





FREE NEWS API

FREE NEWS API

Baza artykułów

TWITTER API

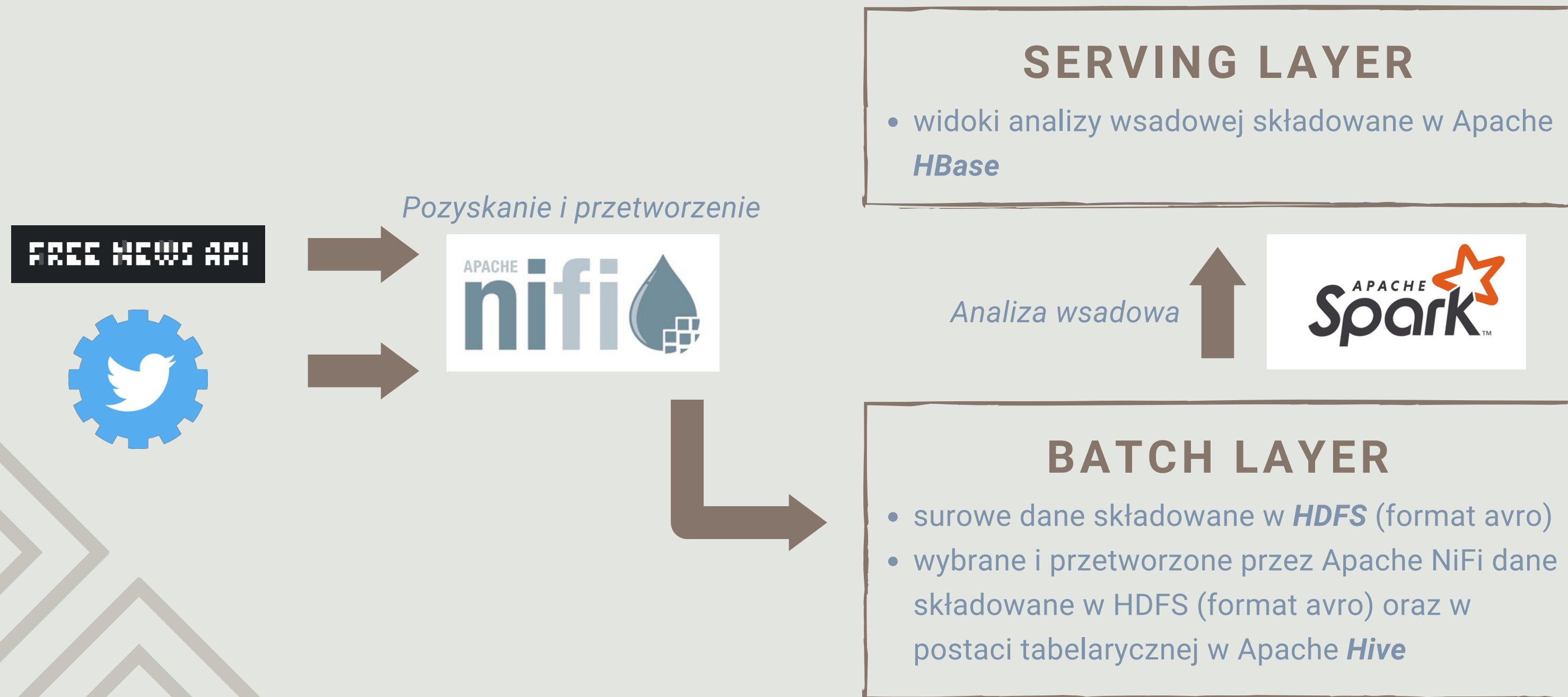
Tweety o artykule

Informacje o wydawnictwach

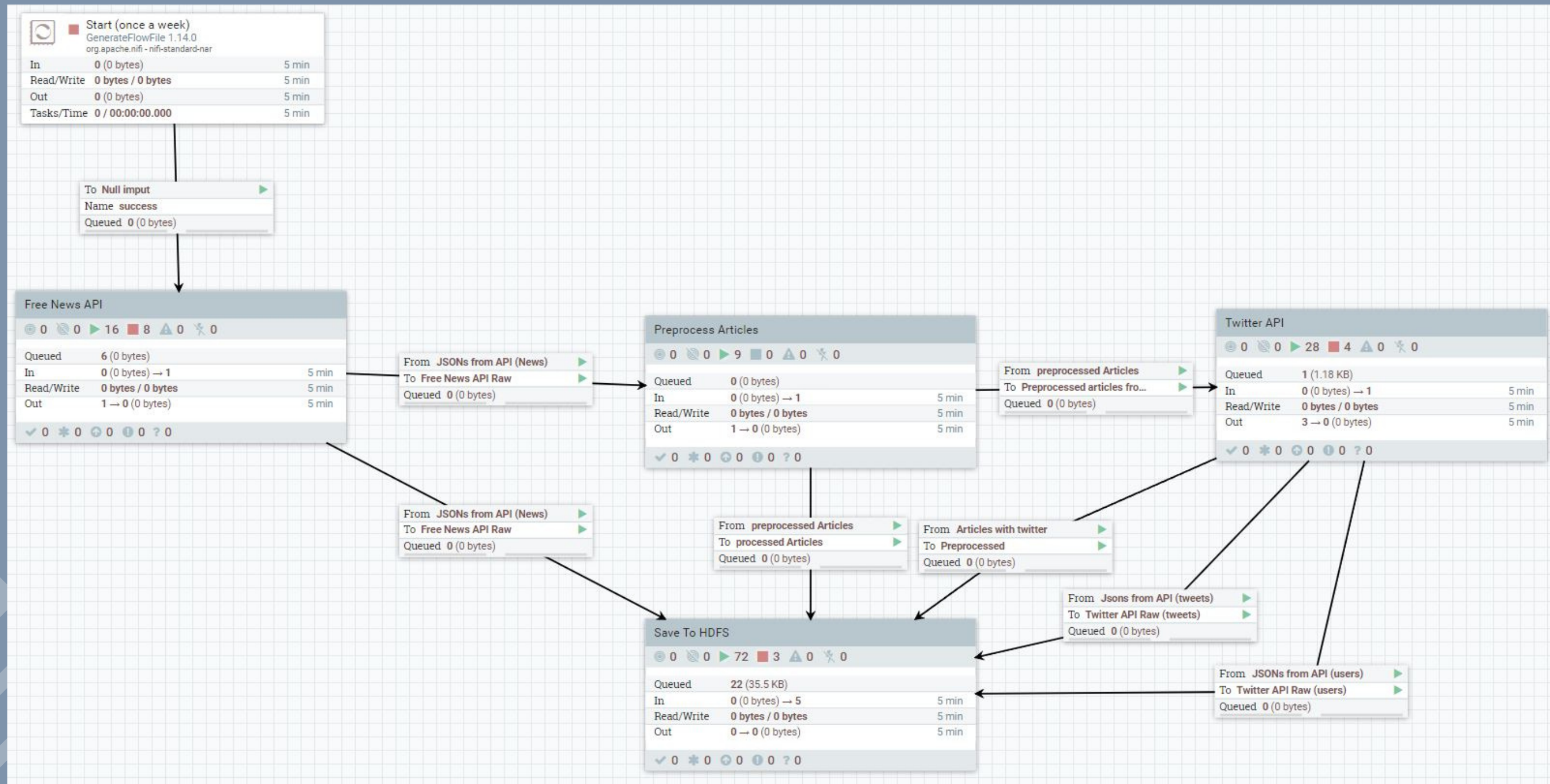


SCHEMAT ROZWIĄZANIA

DIAGRAM ARCHITEKTURY

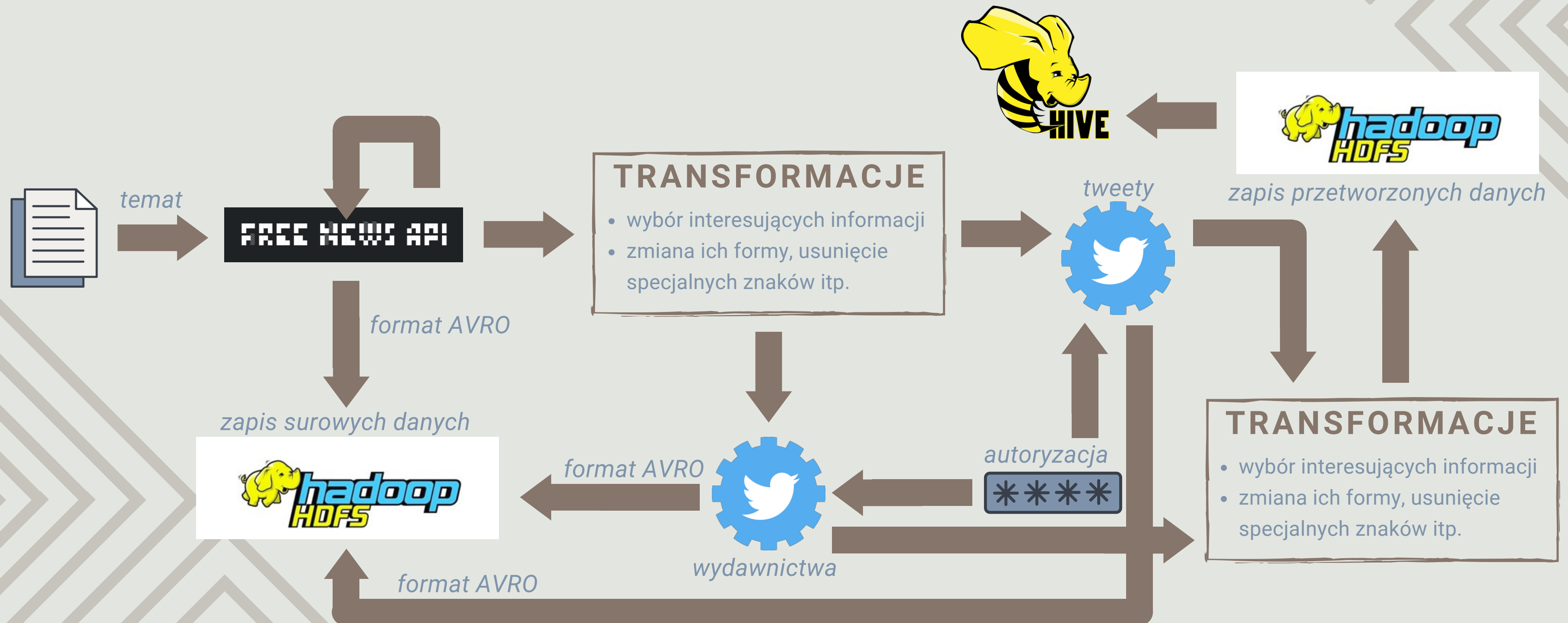


APACHE NIFI, CZYLI POZYSKIWANIE I PRZETWARZANIE DANYCH



POZYSKIWIWANIE I PRZETWARZANIE DANYCH

ÓOGÓLNIONY SCHEMAT DZIAŁANIA



HDFS ORAZ APACHE HIVE, CZYLI SKŁADOWANIE DANYCH (BATCH LAYER)

DANE SUROWE

Dane uzyskane jako odpowiedzi z Twitter API oraz Free News API są z postaci JSON przekształcane na format AVRO i zapisywane w HDFS. Są to pełne dane niezmienione w żaden sposób.

PRZETWORZONE DANE (HDFS)

Tylko część danych jest interesująca i warta wykorzystania do analizy wsadowej, dlatego wybrane informacje z danej odpowiedzi API są przekształcone do odpowiedniej postaci (np. przez usunięcie specjalnych znaków) i zebrane w oddzielny plik w formacie AVRO, następnie zapisany w HDFS.

PRZETWORZONE DANE (HIVE)

Przetworzone dane są zapisywane w HDFS w odpowiednim folderze i od razu podnoszone przez NiFi w celu przepisania ich na postać tabelaryczną i załadowania do istniejących tabel w Hive.

TABELE HIVE

TABELE ARTICLES, PUBLISHERS ORAZ TWEETS

articles

Kolumna	Typ	Opis
id	string	Identyfikator artykułu od Free News API
published_date	string	Data opublikowania artykułu
title	string	Tytuł artykułu
author	string	Nazwisko autora artykułu
topic	string	Temat artykułu (przypisany przez Free News API)
country	string	Kraj którego dotyczy artykuł
language	string	Język, w którym artykuł został opublikowany
is_opinion	boolean	Czy artykuł został oznaczony jako opinia przez wydawcę
query	string	Zapytanie jakiego użyto do pobrania artykułu z FreeNewsAPI
summary	string	Pierwsze 500 znaków artykułu
my_timestamp	bigint	Timestamp

tweets

Kolumna	Typ	Opis
id	string	Identyfikator; połączenie article_id oraz tweet_id
article_id	string	Identyfikator artykułu wspomnianego w tweecie
tweet_id	bigint	Identyfikator dostarczony przez Twittera
tweet_text	string	Przeczyszczony tekst tweeta
my_timestamp	bigint	Timestamp

publishers

Kolumna	Typ	Opis
id	string	Identyfikator
article_id	string	Identyfikator artykułu opublikowanego przez tego wydawcę
twitter_id	bigint	Identyfikator konta wydawcy na Twitterze
twitter_account	string	Nazwa konta wydawcy na Twitterze
publisher_name	string	Nazwa wydawcy
location	string	Lokalizacja wydawcy
followers_count	int	Ilość followersów wydawcy
list_count	int	Ilość list twitterowych, jakie wydawca utworzył na swoim koncie
number_of_tweets	int	Ilość tweetów z konta wydawcy

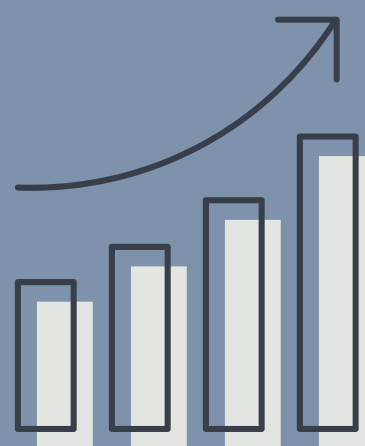
APACHE SPARK, CZYLI ANALIZA WSADOWA KTÓTKO O TYM CO ZROBIONO



Sentyment tytułu artykułu

col1	col2	col3	col4	col5

Tabele agregujące: po
wydawcach lub po tematach



Statystyki: suma artykułów,
procent pozytywnego i
negatywnego sentymentu,
suma tweetów, procent opinii
wśród badanych artykułów

HBASE, CZYLI SKŁADOWANIE WIDOKÓW (SERVING LAYER)

Dwie tabele HBasowe - ***publishers*** oraz ***topics***. Obie tabele mają po trzy column families: name, article_stats oraz twitter_stats:

- name zawiera nazwę tematu lub wydawcy,
- article_stats oraz twitter_stats zawierają odpowiednie statystyki.

```
1 print(topics.head())
```

```
article_stats:articles_with_negative_sentiment_fraction    timestamp \
0                0.0                1642947921065
1                0.341               1642947921598
2                0.357               1642947921631
3                0.2                1642947921270
4                0.274               1642947921199

article_stats:articles_with_positive_sentiment_fraction    article_stats:total_published_articles    name:topic \
0                1.0                1                beauty
1                0.339               528                business
2                0.393               168                economics
3                0.425                40                energy
4                0.46                113               entertainment

twitter_stats:tweets_mentioning_articles_sum
0                0
1                6279
2                619
3                296
4                447
```

Przykładowe wiersze z tabeli topics

PODSUMOWANIE

Co poszło dobrze?

Najtrudniejsze części projektu

Możliwości rozwoju

The background features decorative geometric patterns in the corners. The top-right corner has a series of nested, downward-pointing chevrons. The bottom-left corner has a series of nested, upward-pointing chevrons. Both are composed of multiple parallel lines in a dark brown color.

PYTANIA?

Dziękujemy za uwagę!