WYDZIAŁ MATEMATYKI I NAUK INFORMACYJNYCH
POLITECHNIKA WARSZAWSKA

ANALIZA WPŁYWU CZYNNIKÓW EKONOMICZNYCH NA EKSPORT POSZCZEGÓLNYCH TOWARÓW

SKŁADOWANIE DANYCH W SYSTEMACH BIG DATA

RAPORT

Agata Krawczyk,
Urszula Szczęsna,
Filip Szlingiert

Warszawa, 15 stycznia 2026

Spis treści

1. Wstęp

Celem projektu jest stworzenie systemu do zaciągania, obróbki i składowania danych ekonomicznych, aby umożliwić ich sprawną analizę m.in. wpływu różnych wskaźników makroekonomicznych i kursów walut na eksport poszczególnych dóbr z Polski.

System umożliwia agregowanie danych eksportowych według kraju, rodzaju towaru oraz okresu, co pozwala zidentyfikować czynniki mające największy wpływ na eksport. Analiza pozwoli na ocenę trendów w czasie, zależności między eksportem a sytuacją gospodarczą partnerów handlowych oraz ewentualne przygotowanie rekomendacji dla firm dotyczących rynków docelowych.

Repozytorium projektu zawierające wszystkie pliki niezbędne do odtworzenia rozwiązania znajduje się pod adresem.

2. Zbiory danych

W projekcie zostaną wykorzystane trzy główne źródła danych, które są publicznie dostępne i legalne do przetwarzania w celach analitycznych. Dane będą pobierane przez API lub dostępne w formatach CSV/JSON.

2.1. Dane dotyczące eksportu z Polski

Baza UN Comtrade zawiera dane dotyczące eksportu dla każdego państwa na świecie. W projekcie wykorzystano eksport z Polski w latach 2004 - 2025. Dane będą składowane i odświeżane w ujęciu miesięcznym oraz rocznym.

Dane zawierają:

- kod HS oraz opis towaru
- kod jednostki
- ilość
- waga
- wartość eksportu w USD
- kraj importera
- data importu

Grupy produktów uwzględnionych w analizie:

- 870323 - samochody osobowe
- 871120 - motocykle
- 847130 - sprzęt komputerowy

- 850780 - akumulatory litowo-jonowe/baterie
- 300490 - leki gotowe
- 330499 - kosmetyki
- 100119 - pszenica
- 080810 - jabłka
- 020321 - wieprzowina
- 040690 - sery
- 220830 - wódka
- 240220 - papierosy i wyroby tytoniowe
- 220300 - piwo
- 220290 - napoje bezalkoholowe

2.2. Dane dotyczące kursów walut

API Narodowego Banku Polskiego zawiera zarówno archiwalne, jak i aktualne kursy walut i złota. Dane dotyczące walut dostępne są od 2002 roku i zawierają informację na temat średniego kursu 32 walut w danym dniu roboczym. Odświeżanie planowane jest na koniec każdego miesiąca.

Pozyskano dane dla następujących walut:

- THB – baht tajlandzki (Tajlandia)
- USD – dolar amerykański (Stany Zjednoczone)
- AUD – dolar australijski (Australia)
- HKD – dolar hongkoński (Hongkong)
- CAD – dolar kanadyjski (Kanada)
- NZD – dolar nowozelandzki (Nowa Zelandia)
- SGD – dolar singapurski (Singapur)
- EUR – euro (strefa euro)

- HUF – forint węgierski (Węgry)
- CHF – frank szwajcarski (Szwajcaria)
- GBP – funt szterling (Wielka Brytania)
- UAH – hrywna (Ukraina)
- JPY – jen japoński (Japonia)
- CZK – korona czeska (Czechy)
- DKK – korona duńska (Dania)
- ISK – korona islandzka (Islandia)
- NOK – korona norweska (Norwegia)
- SEK – korona szwedzka (Szwecja)
- RON – lej rumuński (Rumunia)
- TRY – lira turecka (Turcja)
- ILS – nowy szekel izraelski (Izrael)
- CLP – peso chilijskie (Chile)
- PHP – peso filipińskie (Filipiny)
- MXN – peso meksykańskie (Meksyk)
- ZAR – rand (Republika Południowej Afryki)
- BRL – real brazylijski (Brazylia)
- MYR – ringgit malezyjski (Malezja)
- IDR – rupia indonezyjska (Indonezja)
- INR – rupia indyjska (Indie)
- KRW – won południowokoreański (Korea Południowa)
- CNY – juan renminbi (Chiny)
- XDR – specjalne prawa ciągnienia (Międzynarodowy Fundusz Walutowy)

2.3. Dane dotyczące wskaźników makroekonomicznych

Baza World Development Indicators zawiera gamę wskaźników makroekonomicznych wielu państw, grup państw i regionów świata, tj. PKB, populacja, inflacja, wartość importowanych dóbr, wskaźniki rozwoju gospodarczego. Dane zaczęto gromadzić już w 1960 roku, natomiast nie wszystkie wskaźniki są dostępne dla każdego państwa, dlatego do analizy wykorzystano dane z lat 2003-2023.

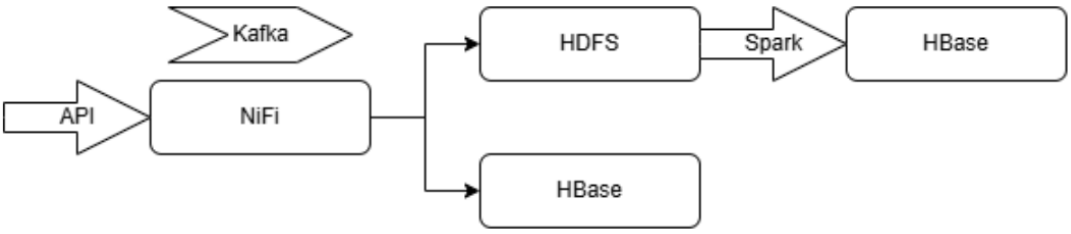
Wskaźniki użyte do analizy to:

- External debt stocks, total (DOD, current US$) - Całkowity dług zagraniczny w dolarach
- GDP (current US$) - PKB w dolarach
- Imports of goods and services (current US$) - Wartość importu w dolarach
- Industry (including construction), value added (% of GDP) - Procentowy udział przemysłu w PKB
- Inflation, consumer prices (annual %) - Inflacja
- Population, total - Liczba ludności
- Services, value added (% of GDP) - Procentowy udział usług w PKB
- Trade (% of GDP) - Procentowy udział handlu w PKB

3. Architektura

Rozwiązanie zostało zaimplementowane w oparciu o narzędzia *open source* przeznaczone do przetwarzania dużych zbiorów danych tj. Apache Hadoop, Apache NiFi, Apache Spark, Apache HBase oraz Apache Kafka. Całość została przygotowana na maszynie wirtualnej przygotowanej przez prowadzących przedmiot.

Architektura rozwiązania nawiązuje do *Lambda Architecture*. Przepływ danych pomiędzy poszczególnymi komponentami systemu przedstawiono na Rysunku 3.1.



Rysunek 3.1: Schemat architektury przepływu danych.

3.1. Apache NiFi

Apache NiFi odegrało kluczową rolę w procesie pozyskiwania danych. Całość danych została pobrana z zewnętrznych interfejsów REST API. Ze względu na zróżnicowaną strukturę źródeł, dla każdego przygotowano oddzielny potok przetwarzania, które następnie zgrupowano w *Process Group*.

Na początku dane zaciągane są przy użyciu zapytań API, następnie część surowych danych jest kierowana do dedykowanych tematów w Apache Kafka. Dane przesłane do Kafki zostają znowu przekazane do NiFi. Ta operacja pozawoliła na oddzielenie procesu pobierania tych danych od ich zapisu. Całość danych jest poddawana wstępnej obróbce i kierowana do systemu

HDFS.

Nietypowym, zastosowaniem Apache NiFi w projekcie było wykorzystanie narzędzia do realizacji symulacji *Speed Layer*. W tym scenariuszu NiFi odpowiadało za bieżące pobieranie strumienia zdarzeń z Apache Kafka, wykonywanie transformacji i agregacji w locie, a następnie bezpośrednie zasilanie tabel w HBase.

3.2. Apache Kafka

Apache Kafka pełni funkcję bufora danych. Jej głównym zadaniem jest zabezpieczenie procesu przetwarzania przed utratą pakietów w przypadku chwilowych problemów z łącznością.

3.3. Apache Hadoop

Rozproszony system plików HDFS stanowi podstawową warstwę składowania dla danych wsadowych. Przechowywane są tam zarówno wstępnie przetworzone zbiory, jak i finalne ramki danych wygenerowane przez Apache Spark.

Strukturę folderów na HDFS:

```
/user/vagrant/project/
|-- comtrade/
|-- final_tables/
|-- NBP/
‘-- WDI/
```

3.4. Apache Spark

Apache Spark umożliwiło docelowe przetwarzanie w warstwie wsadowej. Dane zostały oczyszczone i obsłużono braki, dodano nowe kolumny, łączono zbiory oraz wygenerowano widoki.

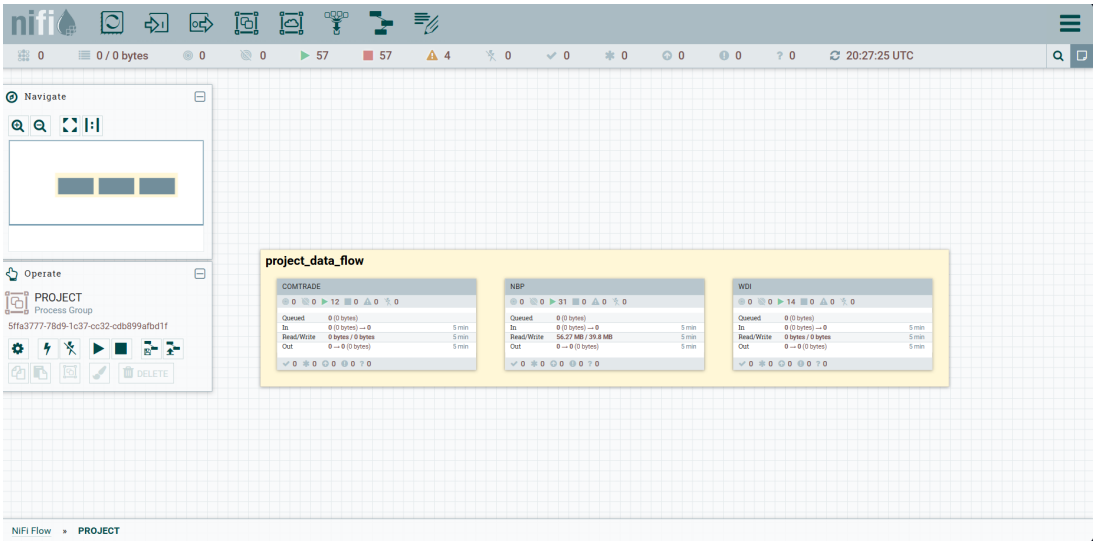
3.5. Apache HBase

W HBase przechowywane były dane z symulaci *Serving Layer* oraz finalne widoki.

Apache HBase pełni rolę bazy danych NoSQL w warstwie *Serving Layer*. Przechowywane są tam wyniki działania warstwy *Speed Layer* oraz zagregowane widoki danych historycznych, umożliwiając szybki dostęp do kluczowych danych.

4. Wstępne przetwarzanie danych

Dane są pobierane za pomocą stworzonych w NiFi przepływów, a następnie wybierane i mapowane są jedynie kolumny niezbędne do dalszych analiz. Przetworzone pliki w formacie *Parquet* i *CSV* są zapisywane w dedykowanych katalogach na HDFS.



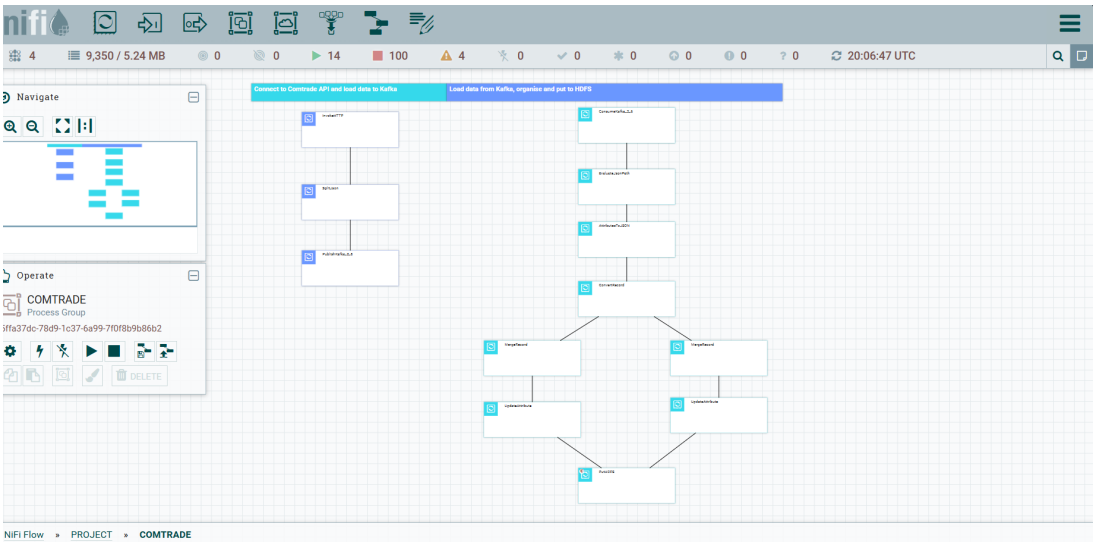
Rysunek 4.1: Widok przepływu z podziałem na zbiory danych.

4.1. Dane UN Comtrade

Po otrzymaniu danych z `ConsumerKafkaRecord` za pomocą `EvaluateJsonPath` sa wybierane i mapowane następujące kolumny:

- cmdDesc - commodity_desc
- period - data_period
- cmdCode - hs_code
- partnerCode - partner_code
- partnerISO - partnerISO
- primaryValue - primary_value_usd
- qty - quantity
- qtyUnitCode - qunatity_code
- netWgt - weight

W przypadku braków danych zapisywana jest wartość *NaN*. Imputacja braków nastąpi dopiero w podczas właściwego przetwarzania w Spark.



Rysunek 4.2: Przepływ danych UN Comtrade.

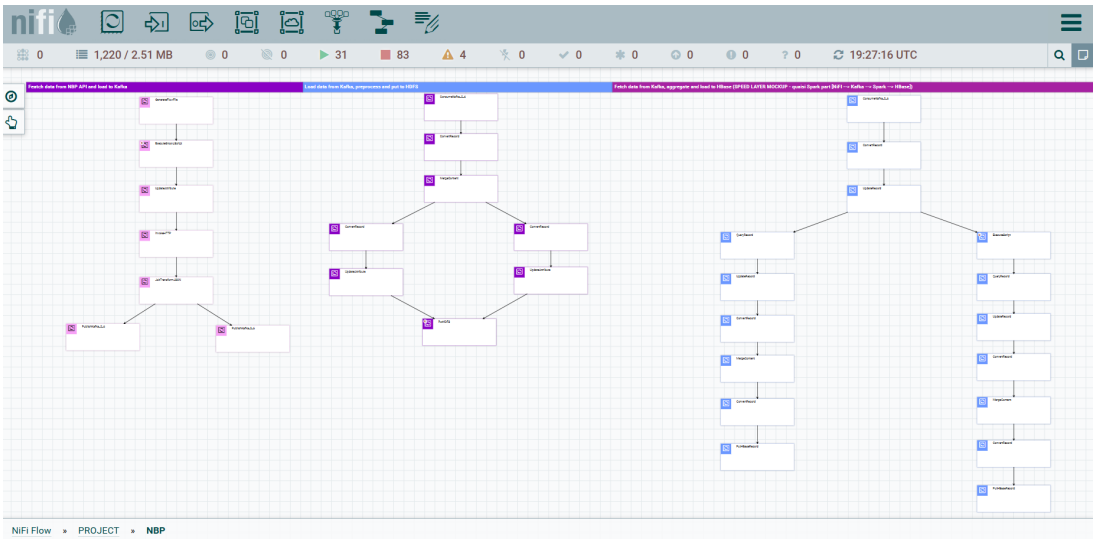
4.1.1. Dane Narodowego Banku Polskiego

Ze względu na limity API dane dla każdego roku i waluty są pobierane oddzielnie, następnie zostają połączone przy użyciu `MergeContent`, poźniej dane ulegają konwersji do farmatu *CSV* oraz *Parquet* i zapisywane są na HDFS.

W danych występują braki dla niektórych walut i lat, w takim przypadku informacje na temat kursu nie zostają włączone do wynikowego zbioru.

Ostatecznie zbiór zawiera następujące kolumny:

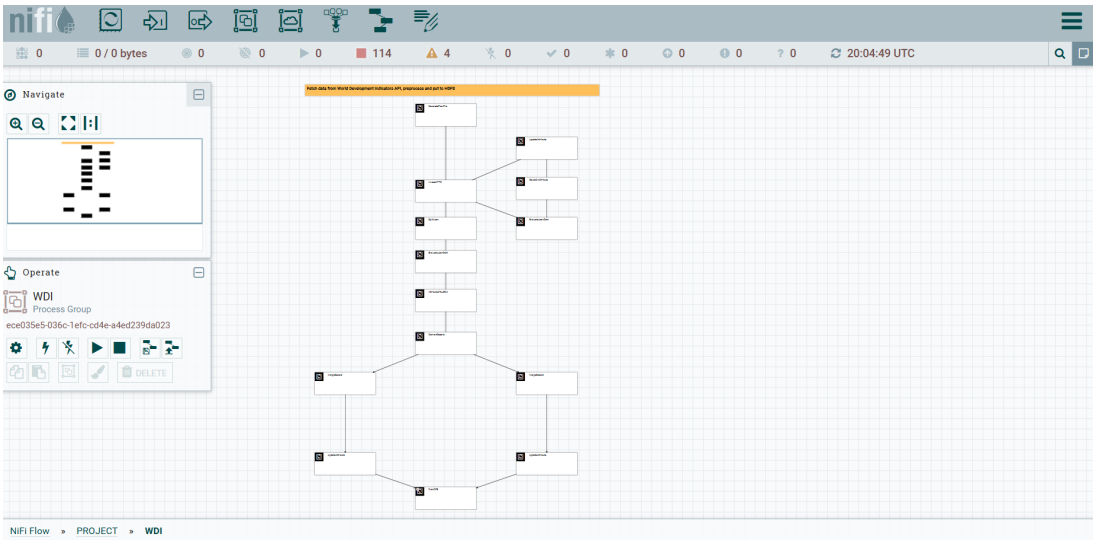
- currency - kod waluty w standardzie ISO 4217
- date - data
- rate - średni kurs waluty z zadanego dnia



Rysunek 4.3: Przepływ danych NBP.

4.2. Dane World Development Indicators

Dane z API spływają partiami. Na początku tworzona jest zmienna `page` identyfikująca numer partii. Za jej pomocą wykonywane jest zapytanie w celu uzyskania odpowiedniej części. Operacja ta dzieje się w pętli a zmienna jest inkrementowana. Zaciągane dane zawierają informację o numerze ostatniej partii, więc wiadomo, kiedy należy przerwać działanie pętli. Spływające dane dzielone są na rekordy, mapowane, a następnie łączone w większe tabele. Ze względu na limit plików mogących oczekiwać w kolejce, dane zapisywane są w kilku mniejszych plikach, zarówno w formacie *Parquet*, jak i *CSV*.



Rysunek 4.4: Przepływ danych NBP.

5. Przetwarzanie danych i tworzenie widoków

Dla każdego ze zbiorów zostały przeprowadzone zaawansowane operacje obejmujące przetwarzanie (np. rozszerzenie o dodatkowe kolumny z kalkulacjami) oraz wstępną analizę danych przy wykorzystaniu Apache Spark w Pythonie. Przygotowano notatnik `spark_code.ipynb`, zawierający kod wszystkich czynności wykonanych na zbiorach.

5.1. Przetwarzanie danych dotyczących eksportu z Polski

Dane wejściowe są ładowane z odpowiedniego folderu na HDFS. Następnie sprawdzane są wartości brakujące (NaN). Ze względu na ich relatywnie niewielką liczbę oraz fakt, że występują wyłącznie w dwóch kolumnach numerycznych (`quantity` i `weight`), zdecydowaliśmy się na uzupełnienie ich wartością 0. Dodatkowo, po wstępnej analizie kodów państw w kolumnie `partnerISO`, zidentyfikowaliśmy niepoprawne kody: *X1 _X*, *E19* oraz *F19* które zostały oznaczone jako *unidentified*.

Tworzone są nowe kolumny:

- **unit_value_usd** – iloraz `primary_value_usd` oraz `quantity`
- **usd_per_kg** – iloraz `primary_value_usd` oraz `weight`
- **year** – rok wyodrębniony z `data_period`
- **month** – miesiąc wyodrębniony z `data_period`
- **quarter** – kwartał wyodrębniony z `data_period`
- **quarter_label** – etykieta kwartału (np. Q1)

Aby obliczenia w obrębie tego samego okresu i towaru umożliwić, zastosowano funkcje okna analitycznego (*window functions*) z partycjonowaniem po kolumnach `data_period`, `commodity_desc` oraz `hs_code`. W ramach każdej partycji wyznaczono miesięczną wartość

światowego eksportu (`month_world_export_value`) jako sumę `primary_value_usd` dla obserwacji z kodem partnera `W00`, a następnie obliczono udział danego kraju w miesięcznym rynku światowym (`share_of_month_market`).

Dane zagregowano do poziomu rocznego według `year`, `partnerISO`, `commodity_desc` oraz `hs_code`, obliczając roczną wartość handlu, łączną ilość oraz wagę. Na tej podstawie wyznaczono wskaźniki jednostkowe (`unit_value_usd`, `usd_per_kg`). Z wykorzystaniem funkcji okna analitycznego obliczono również roczną wartość światowego eksportu (`world_export_value`) oraz udział kraju w rocznym rynku światowym (`share_of_year_market`).

5.2. Przetwarzanie danych dotyczących kursów walut

Dane są ściągane z folderu na HDFS, a następnie powstają dwie ramki danych `yearly_agg` i `monthly_agg`, w których obliczany jest średni, minimalny oraz maksymalny kurs dla każdej waluty w danym roku oraz miesiącu.

Dodatkowo tworzony jest słownik pozwalający na zmapowanie państw do walut, które w nich obowiązują.

5.3. Przetwarzanie danych dotyczących wskaźników makroekonomicznych

Dane ładoowane są z dedykowanego folderu na HDFS. Następnie mapowane są kody państw z notacji iso2 na iso3, która występuje w dwóch pozostałych zbiorach danych. POnad to zmieniane są nazwy poszczególnych wskaźników na krótsze:

- External debt stocks, total (DOD, current US$) - `external_debt`
- GDP (current US$) - `gdp`
- Imports of goods and services (current US$) - `import`
- Industry (including construction), value added (% of GDP) - `industry_in_gdp`
- Inflation, consumer prices (annual %) - `inflation`
- Population, total - `population`
- Services, value added (% of GDP) - `services_in_gdp`

- Trade (% of GDP) - `trade_in_gdp`

Dodatkowo tworzony jest nowy wskaźnik `gdp_per_capita` obliczony jako iloraz `gdp` i `population`.

5.3.1. Połączenie danych

Następnym podjętym krokiem było stworzenie dwóch tabel z połączonymi danymi. Jedna łączy dane z eksportu z Polski oraz dane dotyczące kursów walut w agregacji miesięcznej. Druga zaś łączy wszystkie dane w agregacji rocznej. Decyzja te była umotywowana faktem ,że wskaźniki markoekonomiczne są przedstawiane jedynie w ujęciu rocznym i stosowanie ich w ujęciu miesięcznym byłoby błędem.

```
Schema after all joins:
root
 |-- currency: string (nullable = true)
 |-- year: integer (nullable = true)
 |-- partnerISO: string (nullable = true)
 |-- commodity_desc: string (nullable = true)
 |-- hs_code: string (nullable = true)
 |-- annual_value_usd: double (nullable = true)
 |-- quantity: double (nullable = true)
 |-- weight: double (nullable = true)
 |-- unit_value_usd: double (nullable = true)
 |-- usd_per_kg: double (nullable = true)
 |-- world_export_value: double (nullable = true)
 |-- share_of_year_market: double (nullable = true)
 |-- min_rate: double (nullable = true)
 |-- max_rate: double (nullable = true)
 |-- avg_rate: double (nullable = true)
 |-- country_id: string (nullable = true)
 |-- country_name: string (nullable = true)
 |-- external_debt: double (nullable = true)
 |-- gdp: double (nullable = true)
 |-- import: double (nullable = true)
 |-- industry_in_gdp: double (nullable = true)
 |-- inflation: double (nullable = true)
 |-- population: double (nullable = true)
 |-- services_in_gdp: double (nullable = true)
 |-- trade_in_gdp: double (nullable = true)
 |-- countryISO3: string (nullable = true)
 |-- gdp_per_capita: double (nullable = true)


Created yearly Comtrade with currency rates and WDI (2004-2023)
```

Rysunek 5.1: Schemat połączonych danych rocznych

```
Schema after all joins:
root
 |-- currency: string (nullable = true)
 |-- year: integer (nullable = true)
 |-- month: integer (nullable = true)
 |-- commodity_desc: string (nullable = true)
 |-- data_period: string (nullable = true)
 |-- hs_code: string (nullable = true)
 |-- primary_value_usd: double (nullable = true)
 |-- partner_code: long (nullable = true)
 |-- partnerISO: string (nullable = true)
 |-- quantity: double (nullable = false)
 |-- quantity_code: long (nullable = true)
 |-- weight: double (nullable = false)
 |-- unit_value_usd: double (nullable = true)
 |-- usd_per_kg: double (nullable = true)
 |-- quarter: long (nullable = true)
 |-- quarter_label: string (nullable = true)
 |-- month_world_export_value: double (nullable = true)
 |-- share_of_month_market: double (nullable = true)
 |-- min_rate: double (nullable = true)
 |-- max_rate: double (nullable = true)
 |-- avg_rate: double (nullable = true)


Created monthly Comtrade with currency rates (2004-2023)
```

Rysunek 5.2: Schemat połączonych danych w ujęciu miesięcznym.

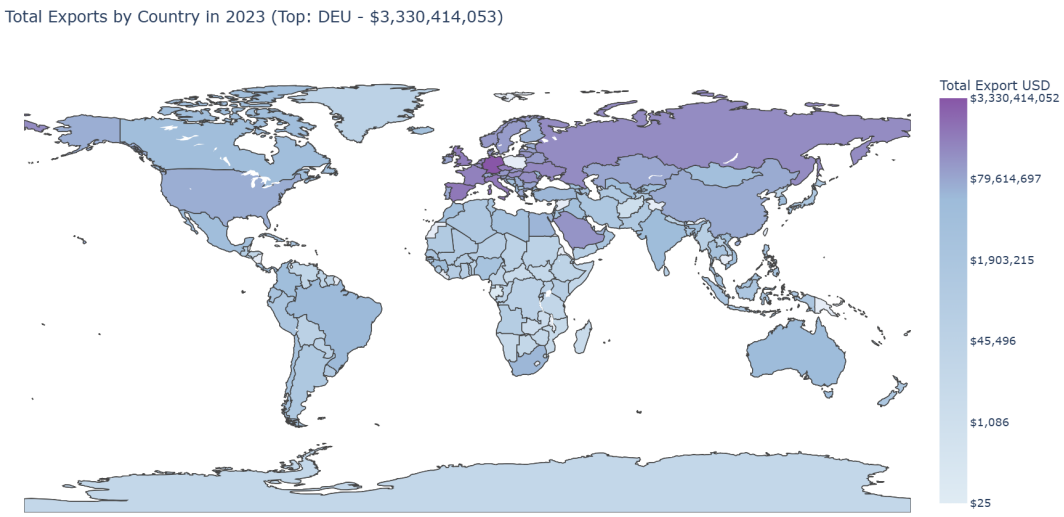
5.4. Widoki

Przygotowano kilka widoków, zapewniających szybki dostęp do sformatowanych danych i będących pomocą podczas analiz:

- Statystyki dotyczące polskiego eksportu
- Struktura gospodarki krajów z największym importem
- Udział w światowym imporcie w 2023 roku
- Ranking dóbr o największej wartości importu

6. Analizy i wizualizacje

6.1. Mapa wartości polskiego eksportu

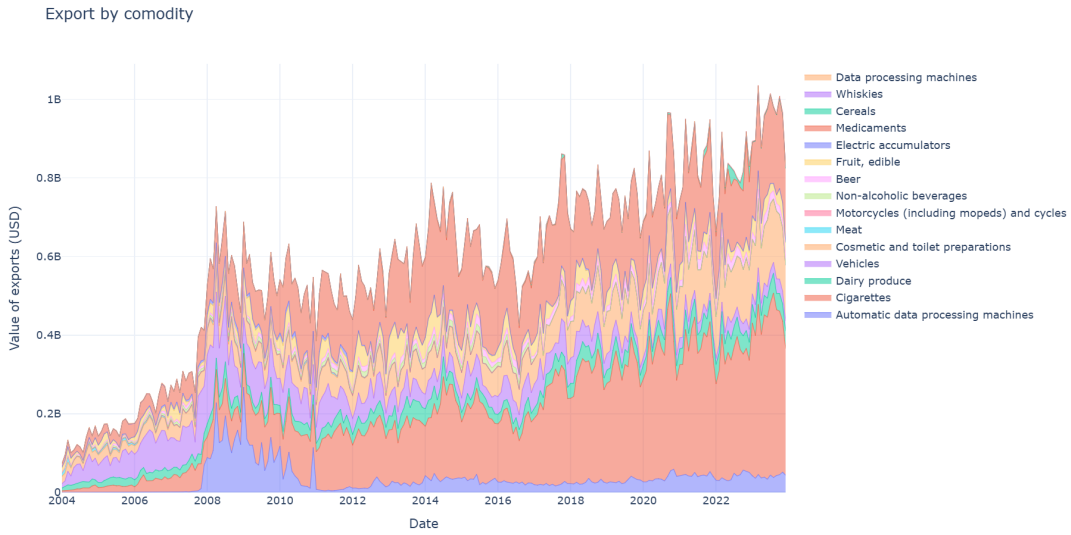


Rysunek 6.1: Mapa cieplna przedstawiająca sumaryczną wartość eksportu.

Wizualizacja przedstawia mapę cieplną z państwami pokolorowanymi zgodnie z sumaryczną wartością towarów importowanych z Polski w 2023 roku. Wyraźnie widać, że najwięcej towarów eksportowane jest do Niemiec i innych państw europejskich. Z mapy wynika, że państwa spoza Europy na na ogół nie importują znacznej ilości dóbr z Polski.

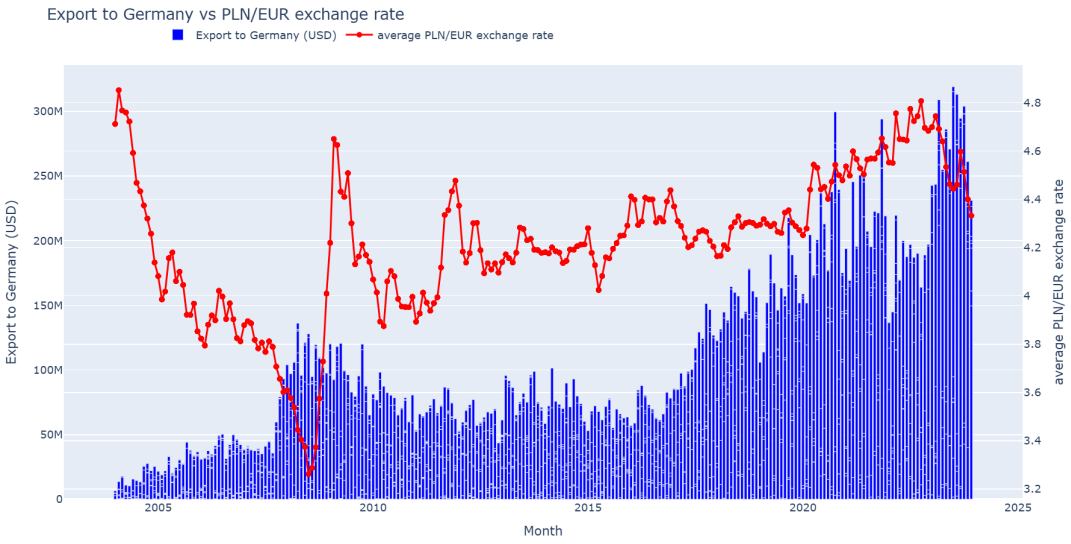
6.2. Wykres wartości polskiego eksportu z podziałem na grupy towarów

Wykres przedstawia jak zmieniały się udział poszczególnych grup towarów eksportowanych z Polski, oraz całość eksportu. Widać na nim zauważalny wzrost eksportu w okolicach roku 2008. Panujący wówczas kryzys ekonomiczny zmusił dostawców do poszukiwania tańszych źródeł towarów. Zdecydowana większość zysków z eksportu pochodzi ze sprzedaży wyrobów medycznych i papierosów.



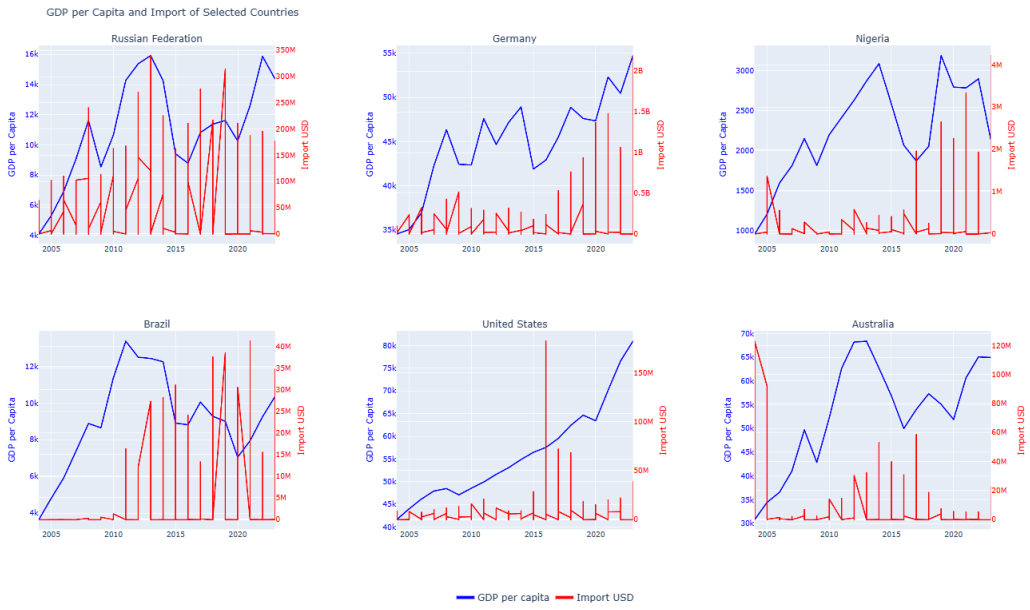
Rysunek 6.2: Wykres udziału poszczególnych towarów w całkowitym eksporcie

6.3. Wykres kursu Euro względem eksportu do Niemiec



Rysunek 6.3: Średni miesięczny kurs euro vs wartość importowanych dóbr.

Wykres przedstawia średni miesięczny kurs euro w porównaniu z wartością importowanych z Polski do Niemiec towarów. Tutaj również widać wyraźny wpływ kryzysu w 2008 roku zarówno na kurs euro jak i wartość importowanych dóbr.



Rysunek 6.4: Wykres PKB per capita i wartości eksportu do wybranych państw

6.4. Wykres PKB per capita i wartości eksportu do wybranych państw

Wykres przedstawia PKB per capita 6 wybranych państw (po jednym z każdego kontynentu) oraz wartość dóbr importowanych do nich z Polski. Można zauważyć, że wartości te są względnie skorelowane, chodź czasem przesunięte w czasie.

6.5. Struktura gospodarki krajów z największym importem

partnerISO	industry_share	services_share	trade_share	trade_value_million
W00	null	null	null	138647.476357192
DEU	25.853604056449225	62.91559988614376	75.94375175649787	25952.014872806005
ITA	21.723291137192735	65.75762561325318	55.68405302869851	9302.437043633001
GBR	18.47957846628564	70.67687419715223	59.645831211397066	8927.094175737
FRA	17.578494535163298	70.18282015300116	61.794952029574226	8607.012980106998
RUS	30.510757160337466	53.72462400294037	49.43784497162305	8284.416835819
NLD	18.651743585004983	68.98956807590353	146.37297561453946	6852.953485204999
CZE	32.22091403914961	56.26368499783521	136.09393213914592	6814.231333383
ESP	22.079508925284298	66.39243571777733	61.53039540887458	6255.662749115002
HUN	25.50627580681497	56.04957224168689	157.7778457478374	5151.169796854999

Rysunek 6.5: Struktura gospodarki krajów z najwyższą łączną wartością importu.

Powyższy widok opisuje strukturę gospodarki krajów w powiązaniu z ich łączną skalą handlu międzynarodowego. Wyniki pokazują wyraźne różnice w modelach gospodarczych i stopniu

otwartości krajów na handel międzynarodowy. Niemcy wyróżniają się jako silnie uprzemysłowiona i bardzo otwarta gospodarka, , podczas gdy Francja, Włochy i Wielka Brytania mają bardziej usługowy i zdywersyfikowany charakter, z istotnym, lecz mniej dominującym udziałem handlu w PKB. Holandia, Czechy i Węgry charakteryzują się ekstremalnie wysokim udziałem handlu w PKB. Rosja natomiast, mimo wysokiego udziału przemysłu, pozostaje gospodarką relatywnie mniej otwartą.

6.6. Udział w światowym imporcie w 2023 roku

year	partnerISO	partner_value_usd	avg_market_share
2023	DEU	3.330414053E9	0.19600777343634954
2023	MMR	118279.0	0.09382623095447805
2023	NLD	4.10097909E8	0.07338411409407901
2023	UKR	3.33064588E8	0.0711358624826894
2023	GBR	5.55879375E8	0.06504885191451125
2023	CZE	5.23948222E8	0.05942135117047254
2023	LVA	1.24665409E8	0.05545737009237996
2023	ITA	6.67133017E8	0.047560868297789244
2023	BLR	1.32979074E8	0.044122062305389495
2023	FRA	4.55614679E8	0.04360252466667384

Rysunek 6.6: Udział w światowym imporcie w 2023 roku.

Powyższy widok prezentuje top 10 państw, które miały największy udział w rynku międzynarodowym w 2023 roku (licząc go jako ułamek wartości importowanych dóbr w porównaniu do całkowitego światowego importu). Dominują tu państwa europejskie, w czołówce są Niemcy, którzy mają wyniki niemal 20% światowego importu.

6.7. Ranking dóbr o największej wartości importu

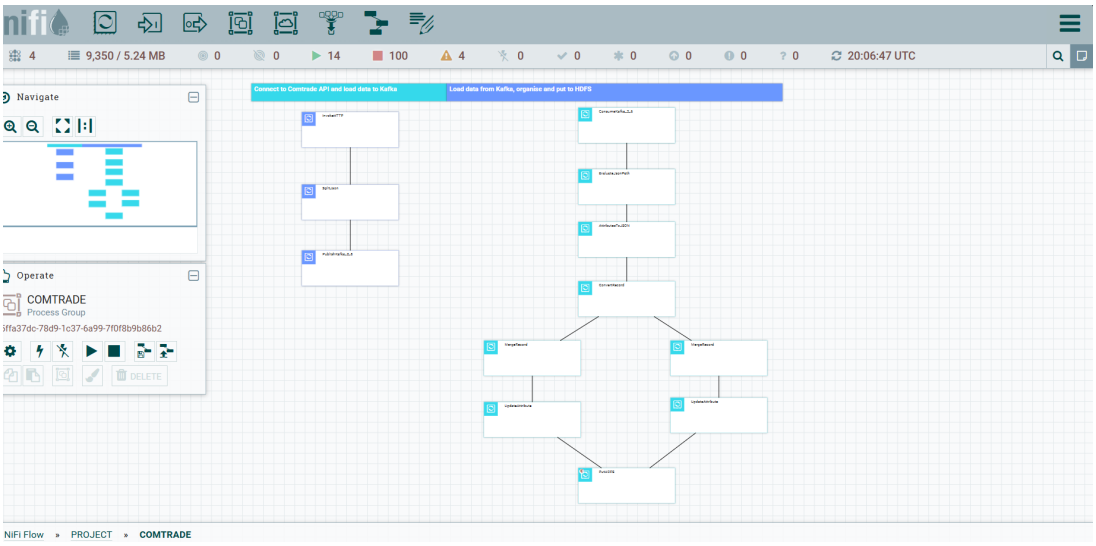
commodity_short	hs_code	total_value_usd	total_quantity	avg_unit_price
Cigarettes	240220	8.5555730968527E10	3.476016126853E9	57.46384337422443
Medicaments	300490	7.10729034747309 9E10	1.0636754523549999E9	209.86171775834967
Cosmetic and toilet preparations	330499	3.5777598236839005E10	3.0654833217660003E9	17.484563479197405
Vehicles	870323	2.7723612507837E10	1725333.23	11552.160134795327
Automatic data processing machines	847130	1.6407660205097E10	3.0935071409999996E7	429.98226274651023
Dairy produce	040690	1.5305122598846996E10	2.3524401801680007E9	3.4156420803494276
Fruit, edible	080810	1.2424106659532001E10	1.7502610449458E10	0.38999991298548015
Beer	220300	5.947720751384999E9	6.727990782E9	0.7235051121437418
Non-alcoholic beverages	220290	2.851283203535E9	1.82614585E9	0.569379265781596
Electric accumulators	850780	1.594059466156E9	2.80598692E8	67.95027930154052
Cereals	100119	1.228087849079E9	2.891581694E9	0.2281083566060143
Meat	020321	9.307613372600001E8	4.3490618E8	1.1859131768663103

Rysunek 6.7: Ranking dóbr o największej sumarycznej wartości importu

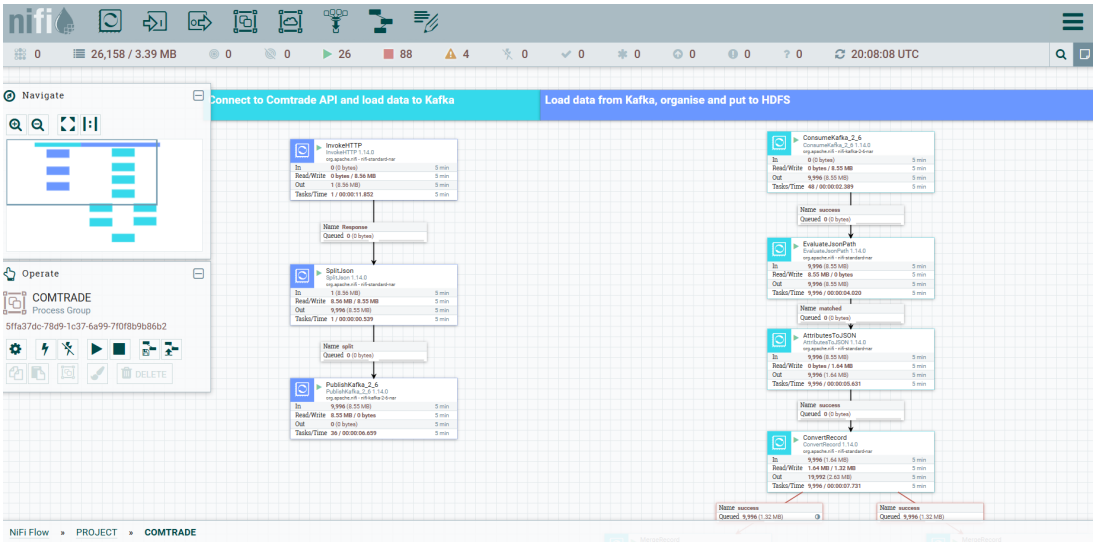
Powyższy widok prezentuje dobra zgodnie z ich sumaryczną wartością importu. Widać że papierosy oraz wyroby medyczne niemal dwukrotnie przewyższają wyroby kosmetyczne czy pojazdy. Jeśli chodzi o produkty spożywcze to wyroby mleczne i owoce dużo przewyższają mięso i zboża.

7. Testy

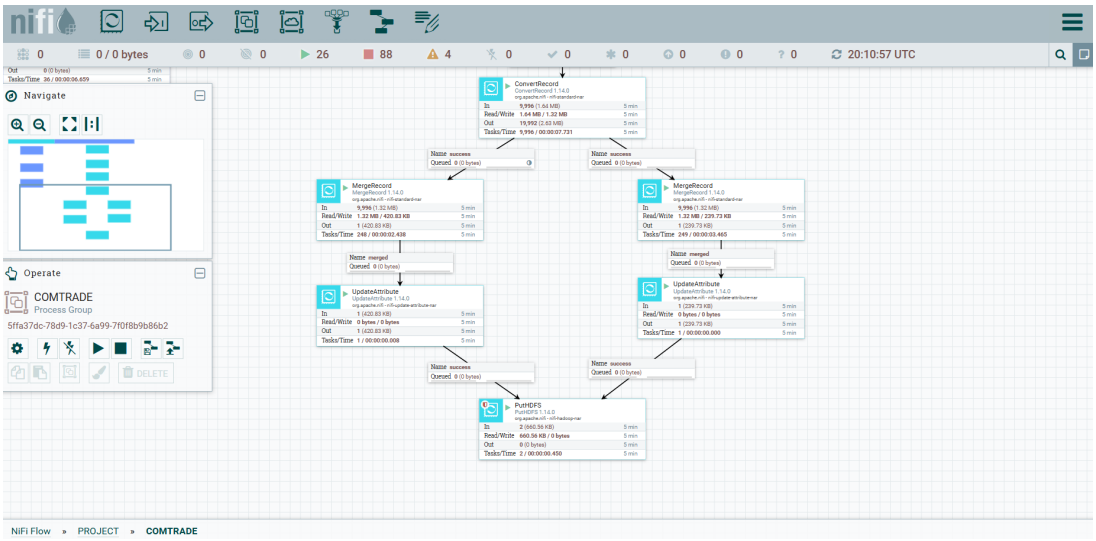
7.1. Przepływ NiFi dla danych Comtrade i poprawność zapisu na HDFS



Rysunek 7.1: Przepływ dla danych UN Comtrade.



Rysunek 7.2: Poprawny zapis i odczyt danych w Kafce.

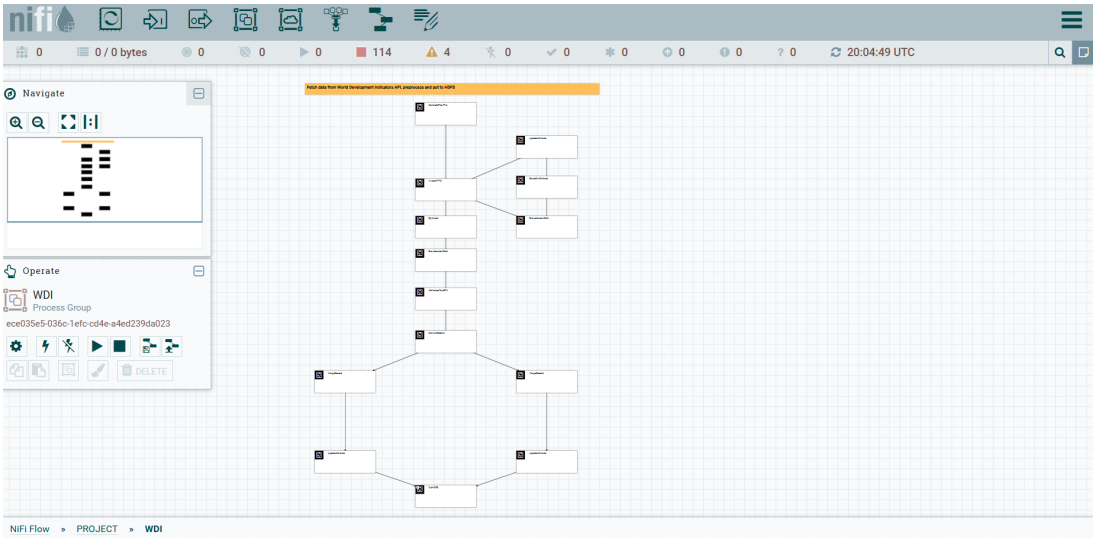


Rysunek 7.3: Poprawny zapis danych w HDFS.

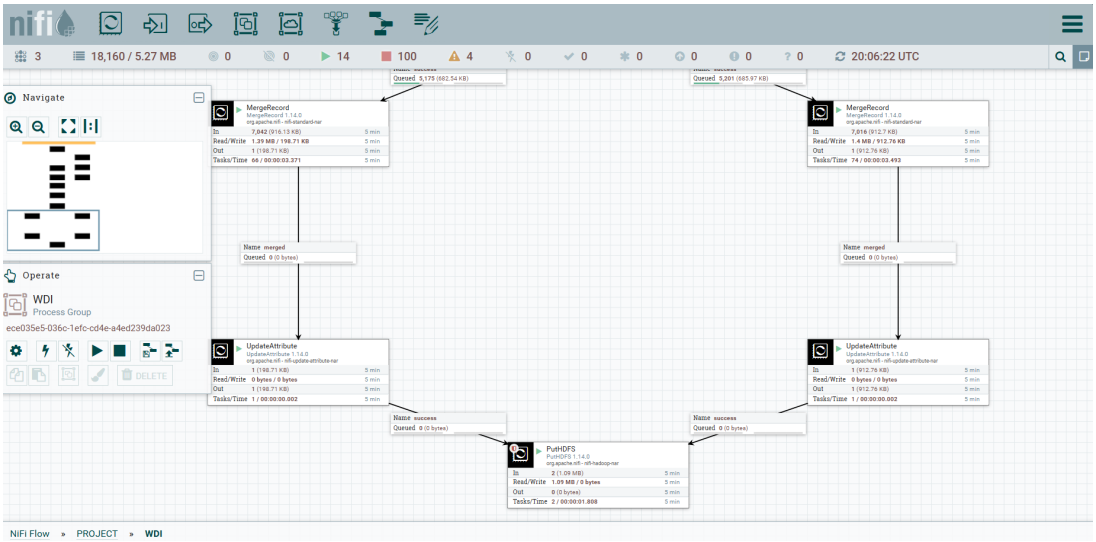
```
vagrant@node1:~$ hadoop fs -ls /user/vagrant/project/comtrade
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/hadoop-2.7.6/share/hadoop/common/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/apache-tez-0.9.1-bin/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
Found 2 items
-rw-r--r--   1 root supergroup     430927 2026-01-14 17:37 /user/vagrant/project/comtrade/comtrade.csv
-rw-r--r--   1 root supergroup     244689 2026-01-14 17:37 /user/vagrant/project/comtrade/comtrade.parquet
vagrant@node1:~$
```

Rysunek 7.4: Sprawdzenie struktury plików w folderze /comtrade.

7.2. Przepływ NiFi dla danych World Development Indicators i poprawność zapisu na HDFS



Rysunek 7.5: Przepływ dla danych WDI.

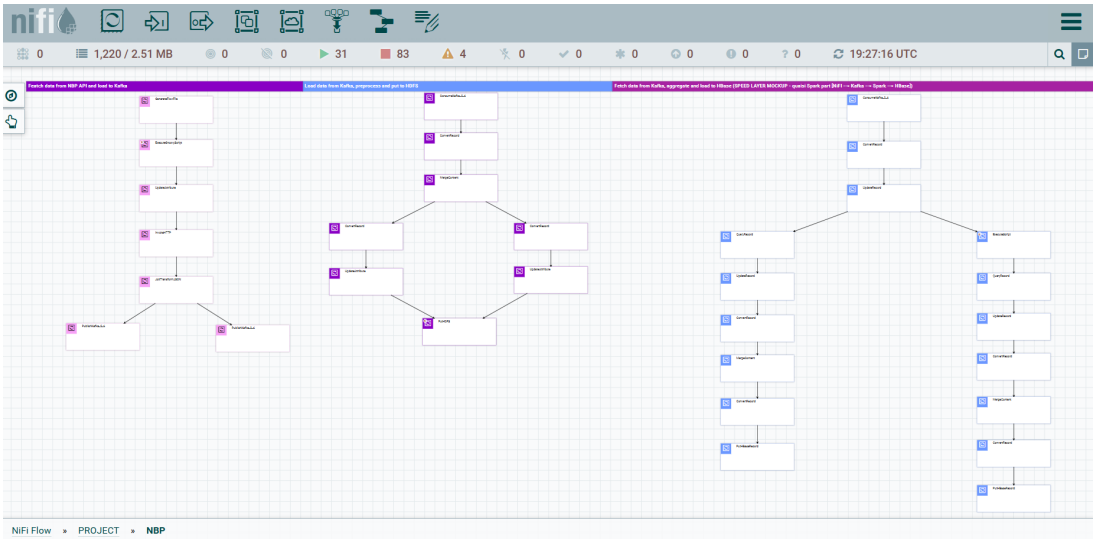


Rysunek 7.6: Poprawny zapis danych w HDFS.

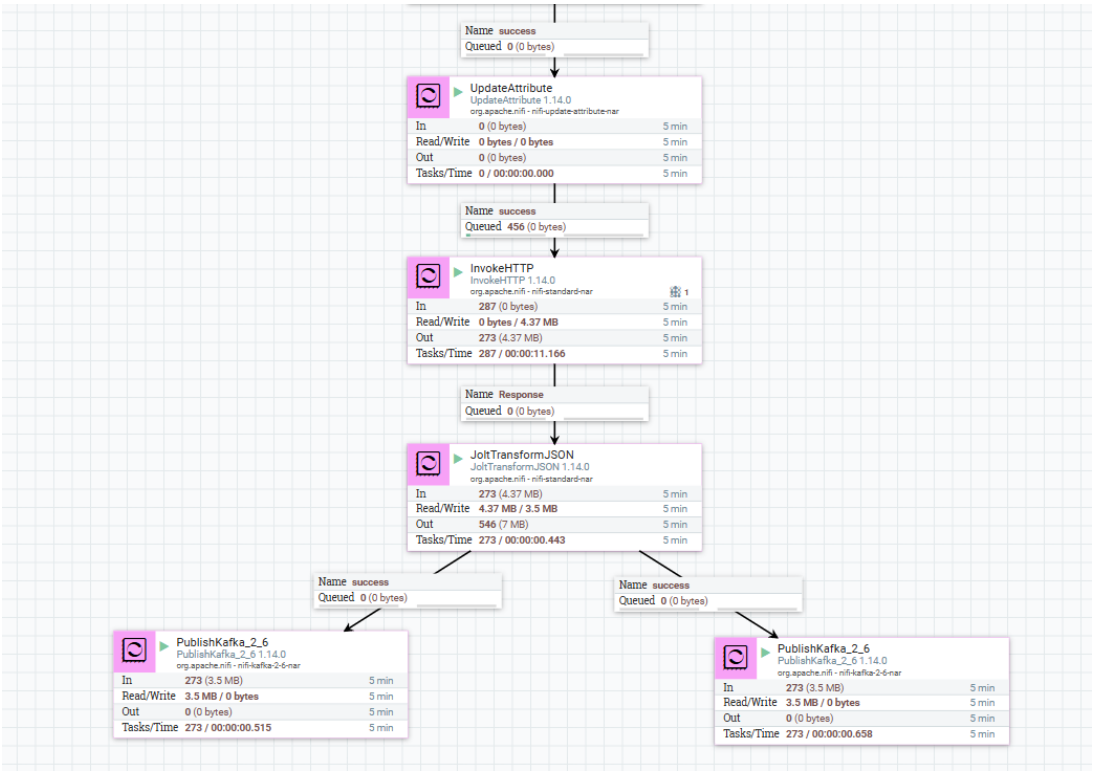
```
vagrant@node1:~$ hadoop fs -ls /user/vagrant/project/WDI
SLF4J: Class path contains multiple SLF4J bindings.
SLF4J: Found binding in [jar:file:/usr/local/hadoop-2.7.6/share/hadoop/common/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: Found binding in [jar:file:/usr/local/apache-tez-0.9.1-bin/lib/slf4j-log4j12-1.7.10.jar!/org/slf4j/impl/StaticLoggerBinder.class]
SLF4J: See http://www.slf4j.org/codes.html#multiple_bindings for an explanation.
SLF4J: Actual binding is of type [org.slf4j.impl.Log4jLoggerFactory]
Found 16 items
-rw-r--r--   1 root supergroup     716285 2026-01-14 17:13 /user/vagrant/project/WDI/wdi_20260114171303936.csv
-rw-r--r--   1 root supergroup     170960 2026-01-14 17:13 /user/vagrant/project/WDI/wdi_20260114171315729.parquet
-rw-r--r--   1 root supergroup     817095 2026-01-14 17:13 /user/vagrant/project/WDI/wdi_20260114171324144.csv
-rw-r--r--   1 root supergroup     133267 2026-01-14 17:13 /user/vagrant/project/WDI/wdi_20260114171324152.parquet
-rw-r--r--   1 root supergroup     125993 2026-01-14 17:13 /user/vagrant/project/WDI/wdi_20260114171330658.parquet
-rw-r--r--   1 root supergroup     701981 2026-01-14 17:13 /user/vagrant/project/WDI/wdi_20260114171333375.csv
-rw-r--r--   1 root supergroup     179800 2026-01-14 17:13 /user/vagrant/project/WDI/wdi_20260114171343624.parquet
-rw-r--r--   1 root supergroup     881965 2026-01-14 17:13 /user/vagrant/project/WDI/wdi_20260114171343809.csv
-rw-r--r--   1 root supergroup     974849 2026-01-14 17:13 /user/vagrant/project/WDI/wdi_20260114171350110.csv
-rw-r--r--   1 root supergroup     147459 2026-01-14 17:13 /user/vagrant/project/WDI/wdi_20260114171350365.parquet
-rw-r--r--   1 root supergroup    1069955 2026-01-14 17:13 /user/vagrant/project/WDI/wdi_20260114171355182.csv
-rw-r--r--   1 root supergroup     173373 2026-01-14 17:13 /user/vagrant/project/WDI/wdi_20260114171356272.parquet
-rw-r--r--   1 root supergroup     197094 2026-01-14 17:14 /user/vagrant/project/WDI/wdi_20260114171400176.parquet
-rw-r--r--   1 root supergroup    1107971 2026-01-14 17:14 /user/vagrant/project/WDI/wdi_20260114171401089.csv
-rw-r--r--   1 root supergroup      84390 2026-01-14 17:16 /user/vagrant/project/WDI/wdi_20260114171601613.parquet
-rw-r--r--   1 root supergroup     375217 2026-01-14 17:16 /user/vagrant/project/WDI/wdi_20260114171602286.csv
vagrant@node1:~$
```

Rysunek 7.7: Sprawdzenie struktury plików w folderze /WDI.

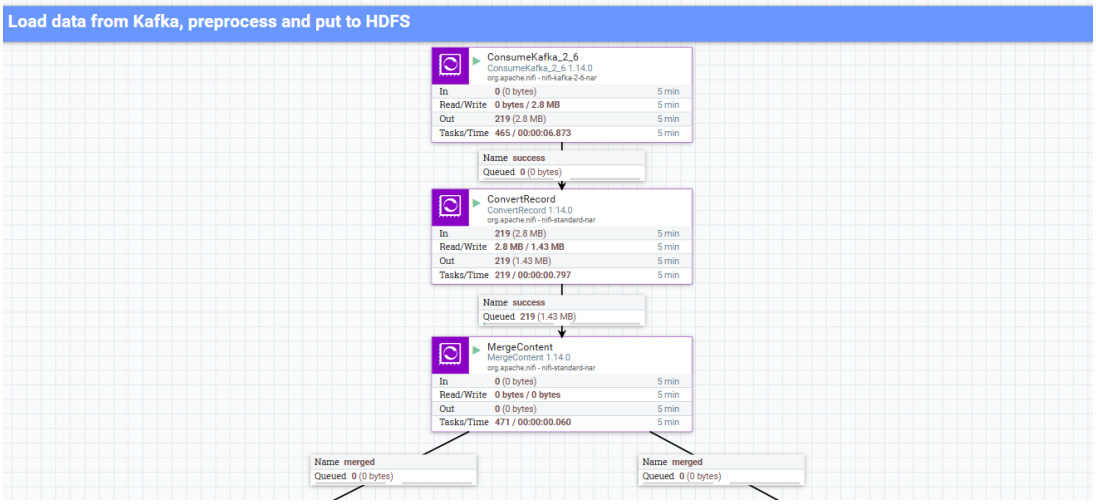
7.3. Przepływ NiFi dla danych Narodowego Banku Polskiego i poprawność zapisu na HDFS



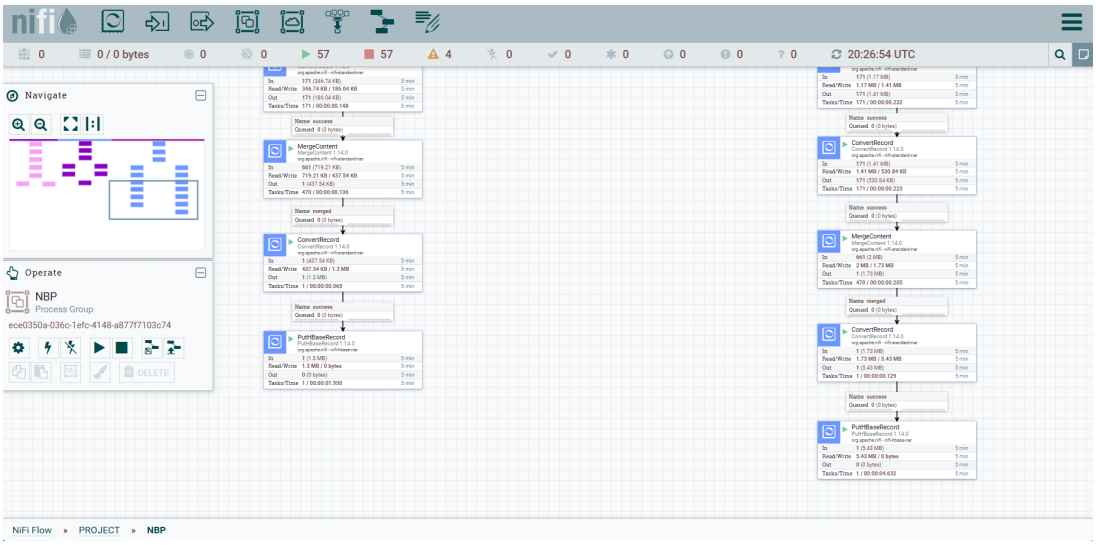
Rysunek 7.8: Przepływ dla danych NBP



Rysunek 7.9: Poprawne umieszczenie danych w tematach Kafka.

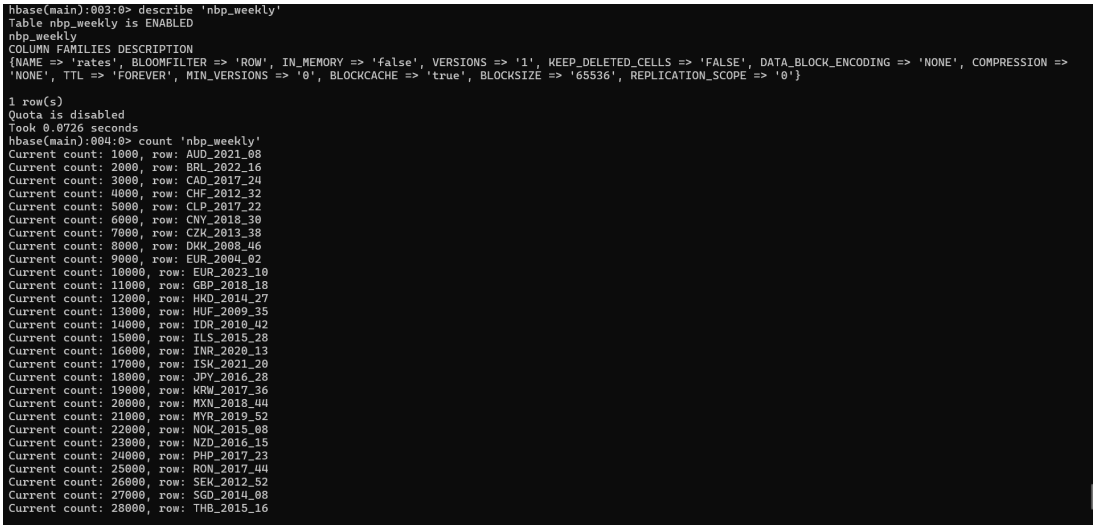


Rysunek 7.10: Poprawne zaciągnięcie danych z Kafka.

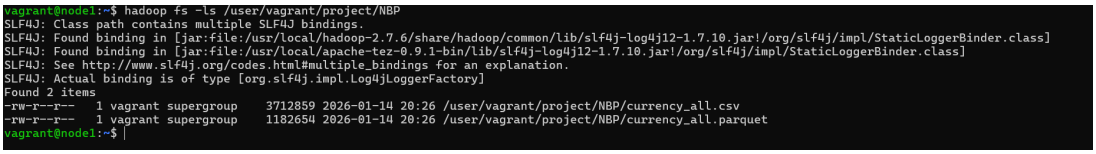


Rysunek 7.11: Poprawny zapis danych do HBase.

Rysunek 7.12: Weryfikacja tabeli miesięcznej w HBase.



Rysunek 7.13: Weryfikacja tabeli tygodniowej w HBase.



Rysunek 7.14: Sprawdzenie struktury plików w folderze /NBP.

7.4. Sprawdzenie zapisu widoków w HBase

Rysunek 7.15: Sprawdzenie listy tabel w HBase.

Kod do sprawdzania tabeli w HBase:

```
def check\_hbase\_table(table\_name):
connection = happybase.Connection('localhost', port=9090)
table = connection.table(table\_name)

for key, data in table.scan(limit=5):
    print(f"Key: {key.decode()}, Data: {data}")

connection.close()
```

Sprawdzenie dla tabeli trade_time

```
Key: 2015\_1, Data: {b'stats:avg\_unit\_price': b'1331.8310130442228', b'stats:month': b'
Key: 2016\_1, Data: {b'stats:avg\_unit\_price': b'1493.3583941708293', b'stats:month': b'
Key: 2017\_1, Data: {b'stats:avg\_unit\_price': b'1508.5282913738733', b'stats:month': b'
Key: 2018\_1, Data: {b'stats:avg\_unit\_price': b'1515.3103545469428', b'stats:month': b'
Key: 2019\_1, Data: {b'stats:avg\_unit\_price': b'1520.6117273497978', b'stats:month': b'
```

Sprawdzenie dla tabeli commodity_ranking

```
Key: 020321\_NO\_NAME, Data: {b'info:avg\_unit\_price': b'2.1036523960069897', b'info:com
Key: 040690\_NO\_NAME, Data: {b'info:avg\_unit\_price': b'5.04279917962624', b'info:commo
Key: 080810\_NO\_NAME, Data: {b'info:avg\_unit\_price': b'0.6116339962682349', b'info:com
```

```
Key: 100119\_NO\_NAME, Data: {b'info:avg\_unit\_price': b'0.4137147752570849', b'info:com
Key: 220290\_NO\_NAME, Data: {b'info:avg\_unit\_price': b'0.8457822076732189', b'info:com
```

Sprawdzenie dla tabeli partner_market_share

```
Key: ABW\_2015, Data: {b'info:avg\_market\_share': b'8.883992663883145e-09', b'info:partn
Key: ABW\_2018, Data: {b'info:avg\_market\_share': b'9.480075055358534e-08', b'info:partn
Key: ABW\_2019, Data: {b'info:avg\_market\_share': b'8.048685960503389e-09', b'info:partn
Key: ABW\_2020, Data: {b'info:avg\_market\_share': b'7.517317341862756e-07', b'info:partn
Key: ABW\_2021, Data: {b'info:avg\_market\_share': b'3.377130616481169e-07', b'info:partn
```

Sprawdzenie dla tabeli economic_structure

```
Key: ABW, Data: {b'info:industry\_share': b'11.947701250438433', b'info:partner\_iso': b'
Key: AFG, Data: {b'info:industry\_share': b'13.145582019192071', b'info:partner\_iso': b'
Key: AGO, Data: {b'info:industry\_share': b'36.86057495976961', b'info:partner\_iso': b'A
Key: AIA, Data: {b'info:industry\_share': b'None', b'info:partner\_iso': b'AIA', b'info:s
Key: ALB, Data: {b'info:industry\_share': b'23.300655186997655', b'info:partner\_iso': b'
```

7.5. Sprawdzenie braków danych

```
=== MISSING VALUES CHECK ===
Total rows: 143342
+-------------+-----------+-------+-----------------+------------+----------+--------+-------------+------+
|commodity_desc|data_period|hs_code|primary_value_usd|partner_code|partnerISO|quantity|quantity_code|weight|
+-------------+-----------+-------+-----------------+------------+----------+--------+-------------+------+
|0            |0          |0      |0                |0           |0         |7386    |0            |1075  |
+-------------+-----------+-------+-----------------+------------+----------+--------+-------------+------+
```

Rysunek 7.16: Braki danych w danych odnośnie eksportu które potem zostały zastąpione 0

```
Null dates check:

+--------+----+----+----+-----+
|currency|date|rate|year|month|
+--------+----+----+----+-----+
+--------+----+----+----+-----+
```

Rysunek 7.17: Raport odnośnie braków danych w danych odnośnie walut

7.6. Weryfikacja mapowania kodów państw

```
=== MAPPING ISO CODES ===

+----------+-----------+
|country_id|countryISO3|
+----------+-----------+
|        SZ|        SWZ|
|        TL|        TLS|
|        KI|        KIR|
|        PT|        PRT|
|        ES|        ESP|
+----------+-----------+
only showing top 5 rows
```

Rysunek 7.18: mapowanie kodów państw

8. Podział pracy

	Agata	Urszula	Filip
Idea projektu	X		
Pozyskiwanie danych	X	X	X
Przygotowanie przepływów danych			X
Składowanie danych		X	
Przetwarzanie danych	X		X
Wsadowa analiza danych		X	
Przygotowanie wizualizacji	X		X
Sporządzenie dokumentacji	X	X	X
Utrzymanie projektu na GitHubie	X	X	X

Rysunek 8.1: Podział obowiązków w projekcie