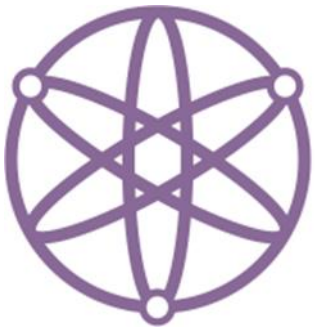


WYDZIAŁ
MATEMATYKI
I FIZYKI STOSOWANEJ
POLITECHNIKI RZESZOWSKIEJ

Wielowymiarowa analiza danych

Projekt

Gabriel Lichacz
Patryk Motyka

Rzeszów, 2022

Spis treści

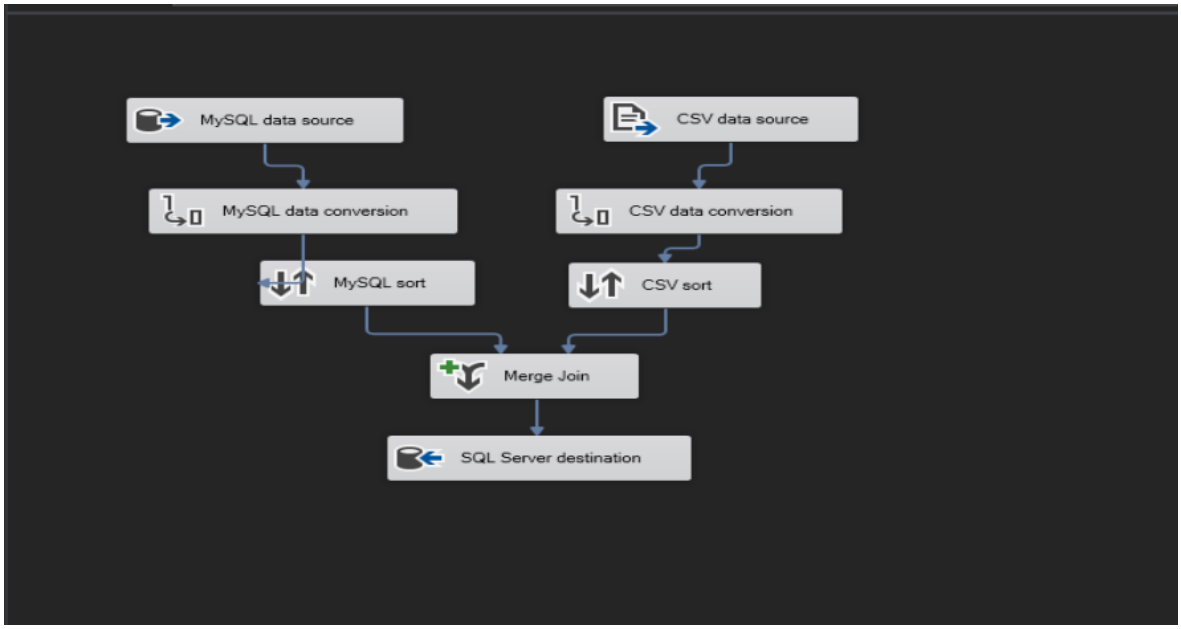
1. Wstęp

Projekt wymagał od nas przejścia kilku etapów, pierwszym z nich było znalezienie danych, które później przeszły proces ETL (ang. Extract, Transform and Load). Zaczęliśmy od pracy w SQL Server Integration Services, później podobne kroki wykonywaliśmy także za pomocą skryptu w języku Python. Dane przechowywaliśmy w magazynie Apache Druid, aby potem umieścić je w baize MySQL. Następnie za pomocą Knime dokonaliśmy procesu ETL stricte pod wizualizację, a także tworzyliśmy kostki Rolap. W ostatniej fazie wykorzystaliśmy klastrowanie i SVM, czyli algorytmy data mining.

2. ETL – SSIS

SQL Server Integration Services to platforma do integracji i transformacji danych. Dzięki usługom integracji pomaga rozwiązać złożone problem biznesowe związane z ładowaniem magazynów danych, czyszczeniem i eksploracją danych oraz zarządzaniem obiektami i danymi program SQL Server.

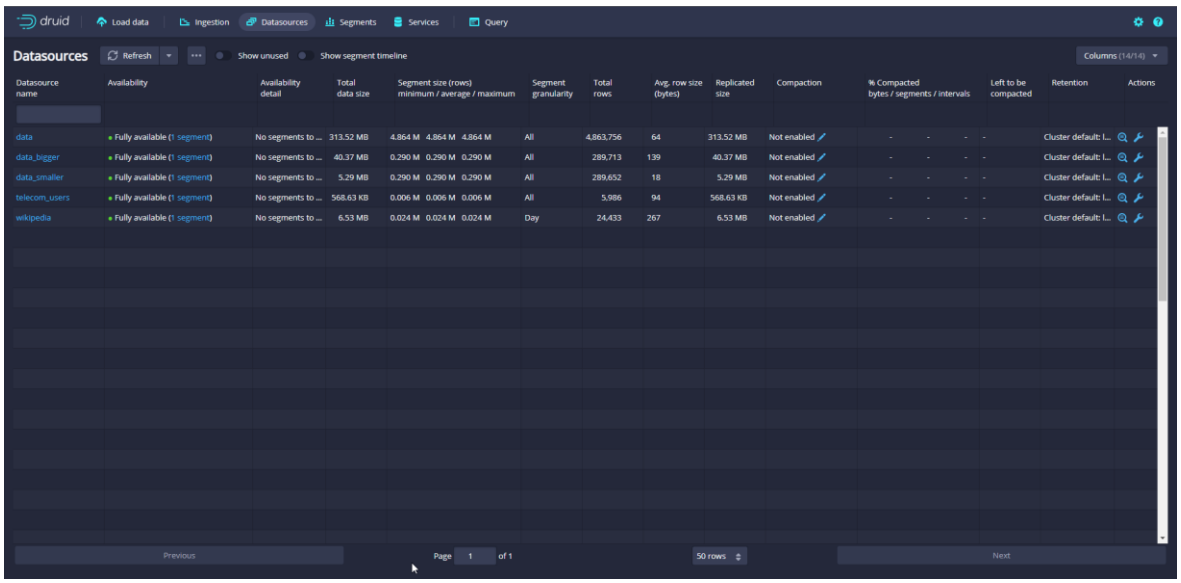
Dane wczytywaliśmy z bazy MySQL z pomocą zewnętrznych sterowników do SSIS. Bazą docelową była baza SQL Server.



2-1 Przepływ danych, który mieliśmy przed problemami z maszyną

3. ETL – Python

Z racji na problem z maszyną wirtualną z narzędziem SSIS postanowiliśmy, że wykonamy ETL poprzez Python z wykorzystaniem bibliotek pydruid oraz sqlalchemy. Dane pobieraliśmy z dwóch tabel w Apache Druid oraz pliku CSV.**ETL wykonany w Pythonie odpowiada przedstawionemu powyżej schematowi w SSIS.**



3-1 Tabele z danymi w Druid

3.1. Skrypt

Import bibliotek

```
# Podstawa
import numpy as np
import pandas as pd
# SQLAlchemy
from sqlalchemy import create_engine
# Kwerendy druid
from pydruid import *
from pydruid.client import *
from pylab import plt
from pydruid.query import QueryBuilder
from pydruid.utils.postaggregator import *
from pydruid.utils.aggregators import *
from pydruid.utils.filters import *
```

Polaczenie z bazą danych w Apache Druid

```
query = PyDruid('http://localhost:8888', 'druid/v2/')
```

Kwerendy wczytujące dane

```
# Wczytanie calej tabeli data_smaller
data_smaller = query.scan(
            datasource = "data_smaller",
            granularity = 'all',
            intervals       =       "-146136543-09-08T08:23:32.096Z/146140482-04-
24T15:36:27.903Z"
)
df_data_smaller = query.export_pandas() # Zamiana danych w ramke danych pandas
print (df_data_smaller)
```

```
# Wczytanie calej tabeli data_bigger
data_bigger = query.scan(
            datasource = "data_bigger",
            granularity = 'all',
            intervals      =      "-146136543-09-08T08:23:32.096Z/146140482-04-
24T15:36:27.903Z"
)
df_data_bigger = query.export_pandas() # Zamiana danych w ramke danych pandas
print (df_data_bigger)
```

Czyszczenie danych

```
# Usuwanie niepotrzebnych kolumn
df_data_smaller = df_data_smaller.drop(['__time'], axis = 1)
df_data_bigger = df_data_bigger.drop(['__time'], axis = 1)

# Sortowanie kolumn po id
df_data_smaller = df_data_smaller.sort_values('id')
df_data_bigger = df_data_bigger.sort_values('id')

# Wyrownanie liczby wierszy
df_data_bigger = df_data_bigger.drop(labels = range(0,61), axis = 0)

# Usuwanie niepotrzbnej juz kolumny
df_data_smaller = df_data_smaller.drop(['id'], axis = 1)

# Polaczenie ramek danych
df_data = pd.concat([df_data_bigger, df_data_smaller], axis = 1)

# Sprawdzenie formatu kolumn
df_data.info()

# Konwersja formatow kolumn
cols = ['id', 'year', 'comm_code', 'trade_usd', 'weight_kg', 'quantity', 'IMO']
df_data[cols] = df_data[cols].apply(pd.to_numeric, errors = 'coerce')
df_data.info()

# Zamiana wartosci NA na puste
df_data['country_or_area'] = df_data['country_or_area'].str.replace('NA','')
df_data['commodity'] = df_data['commodity'].str.replace('NA','')
df_data['flow'] = df_data['flow'].str.replace('NA','')
df_data['category'] = df_data['category'].str.replace('NA','')
df_data['SHIP.NAME'] = df_data['SHIP.NAME'].str.replace('NA','')
df_data['CLASS'] = df_data['CLASS'].str.replace('NA','')
df_data['STATUS'] = df_data['STATUS'].str.replace('NA','')
df_data['REASON.FOR.THE.STATUS']                                                =
df_data['REASON.FOR.THE.STATUS'].str.replace('NA','')
```

Wczytanie danych z pliku CSV

```
df_data_mega = pd.read_csv('/home/gabriel/Desktop/data_wieksze.csv')
```

Czyszczenie danych

```
# Usuwanie niepotrzenych kolumn
df_data_mega = df_data_mega.drop(['quantity_name'], axis = 1)

# Usuwanie zbyt duzej ilosci kolumn (dla szybkosci dzialania MySQLa)
df_data_mega = df_data_mega.drop(labels = range(700000,1436218), axis = 0)

# Konwersja kolumn na numeric
df_data_mega.info()
cols = ['id', 'year', 'comm_code', 'trade_usd', 'weight_kg', 'quantity', 'IMO']
df_data_mega[cols] = df_data_mega[cols].apply(pd.to_numeric, errors = 'coerce')
df_data_mega.info()

# Zamiana wartosci NA na puste
df_data_mega['country_or_area']                                                 =
df_data_mega['country_or_area'].str.replace('nan','')
df_data_mega['commodity'] = df_data_mega['commodity'].str.replace('nan','')
df_data_mega['flow'] = df_data_mega['flow'].str.replace('nan','')
df_data_mega['category'] = df_data_mega['category'].str.replace('nan','')
df_data_mega['SHIP.NAME'] = df_data_mega['SHIP.NAME'].str.replace('nan','')
df_data_mega['CLASS'] = df_data_mega['CLASS'].str.replace('nan','')
df_data_mega['STATUS'] = df_data_mega['STATUS'].str.replace('nan','')
df_data_mega['REASON.FOR.THE.STATUS']                                           =
df_data_mega['REASON.FOR.THE.STATUS'].str.replace('nan','')

# Polaczenie dwoch ramek danych
df = pd.concat([df_data_mega, df_data])
```

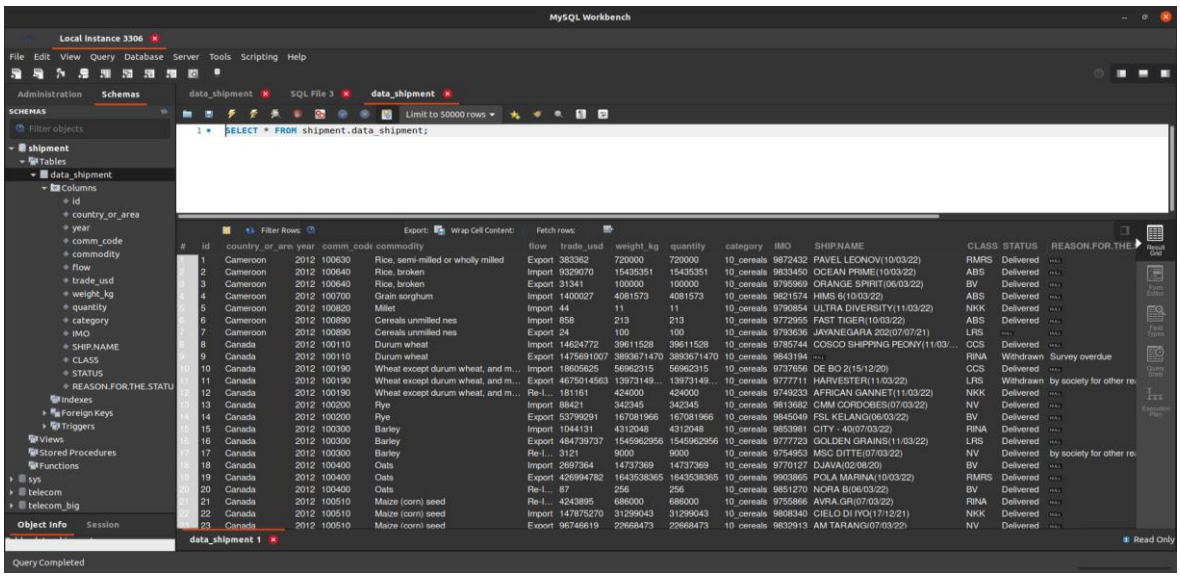
Zapis danych do bazy w MySQL

```
# Polaczenie z baza mysql
# login, haslo, host, baza danych
engine                                                                          =
create_engine('mysql+pymysql://mysql_admin:mysql_admin@localhost:3306/shipment'
)

# Usuwanie tabeli przed wczytaniem
engine.execute('DROP TABLE data_shipment;')

# Zapis do nowej tabeli data_shipment
df.to_sql('data_shipment', engine, index = False)

# Zamiana kolumny id na primary key
#engine.execute('ALTER TABLE data_shipment ADD PRIMARY KEY (`id`);')
```



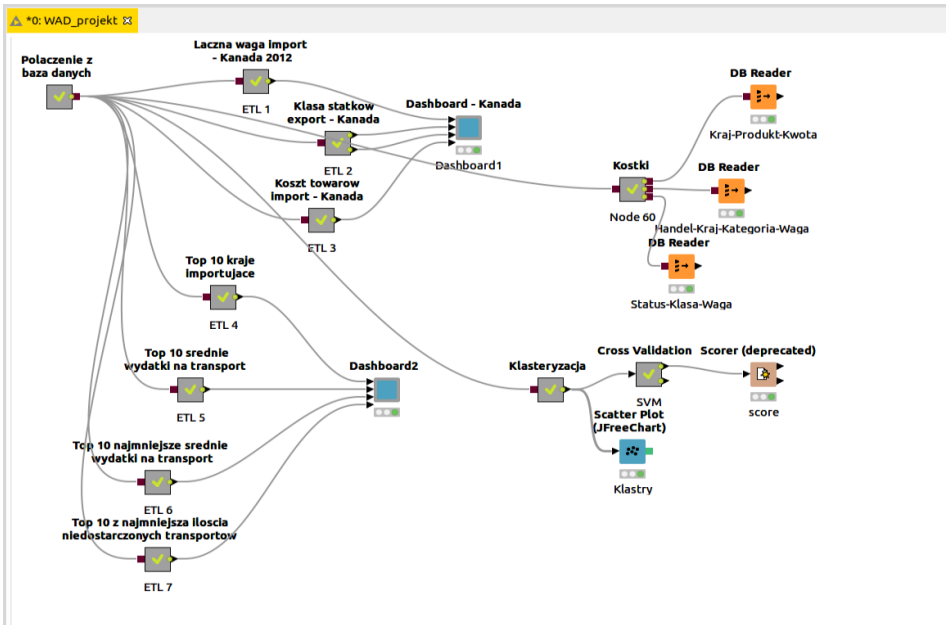
3-2 Dane w MySQL po wykonanym procesie ETL

4. Knime

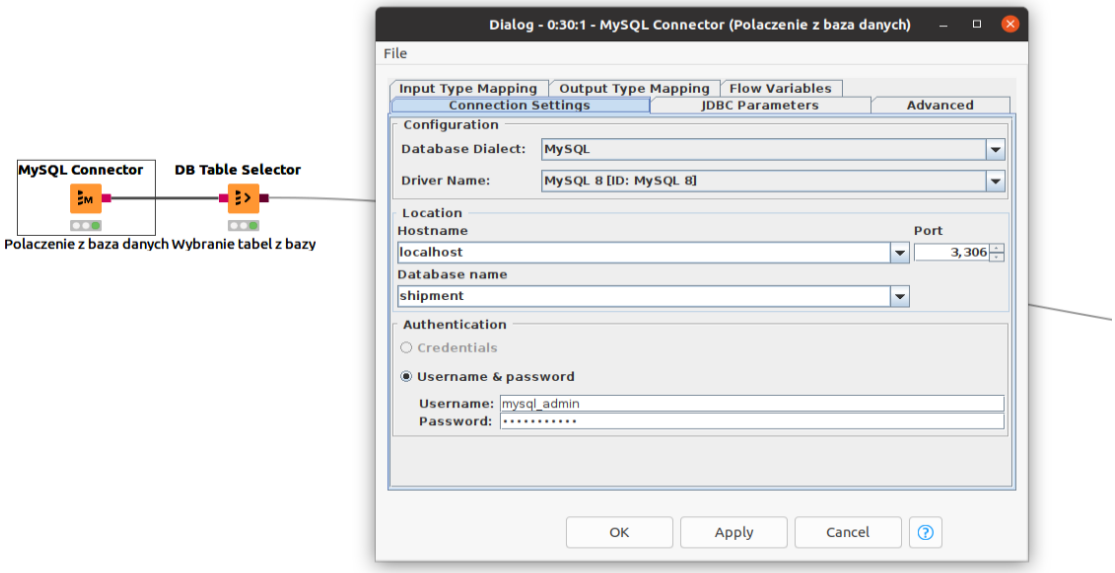
Knime to platforma przeznaczona do integracji oraz analizy, a także raportowania danych. Operacje przeprowadza się za pomocą zestawiania węzłów (nodes), w których można także korzystać z języków jak np. SQL. Knime jest dostępny na zasadzie otwartego oprogramowania.

4.1. ETL

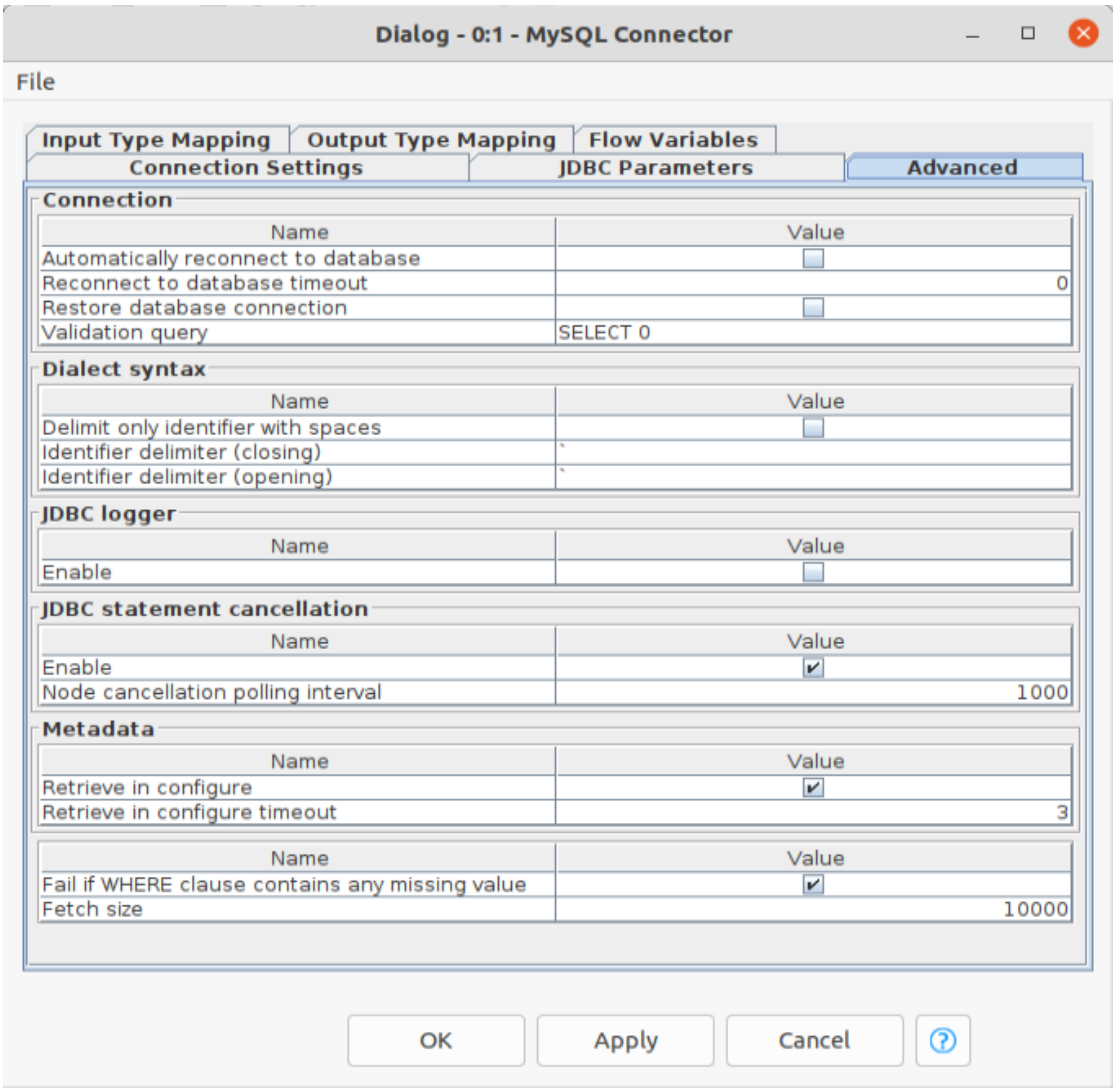
Na pierwszym screenie widoczny jest cały workspace, z ‘metanodes’. Metanodes czyli szare elementy z zielonym symbolem to fragmenty kodu (gotowe węzły zdefiniowane, dostępne w Knime) połączone i zamknięte razem, aby na głównym ekranie zachować większą przejrzystość i zmniejszyć liczbę elementów widocznych na pierwszy rzut oka kilkukrotnie.



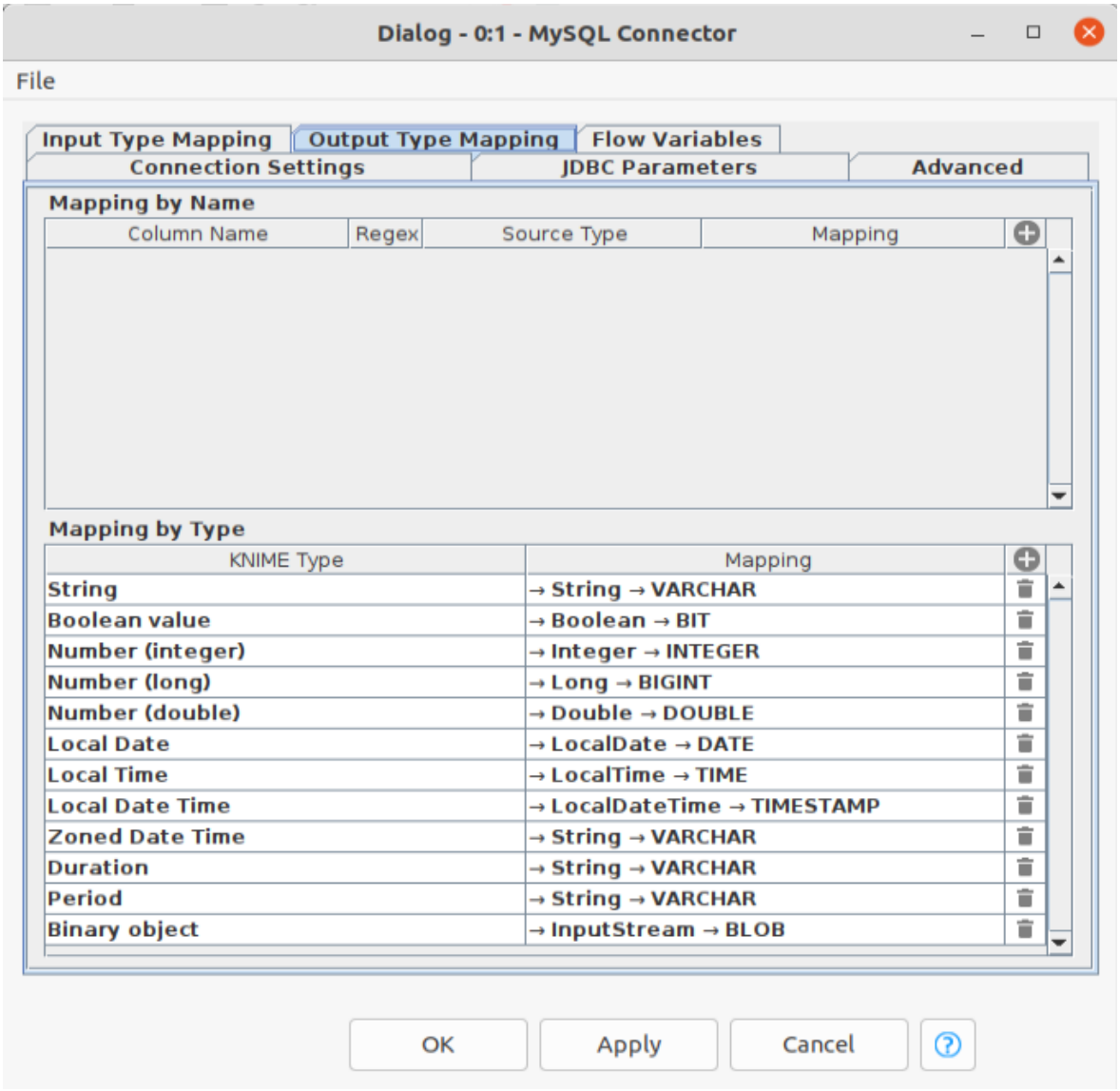
4-1 Workspace w Knime



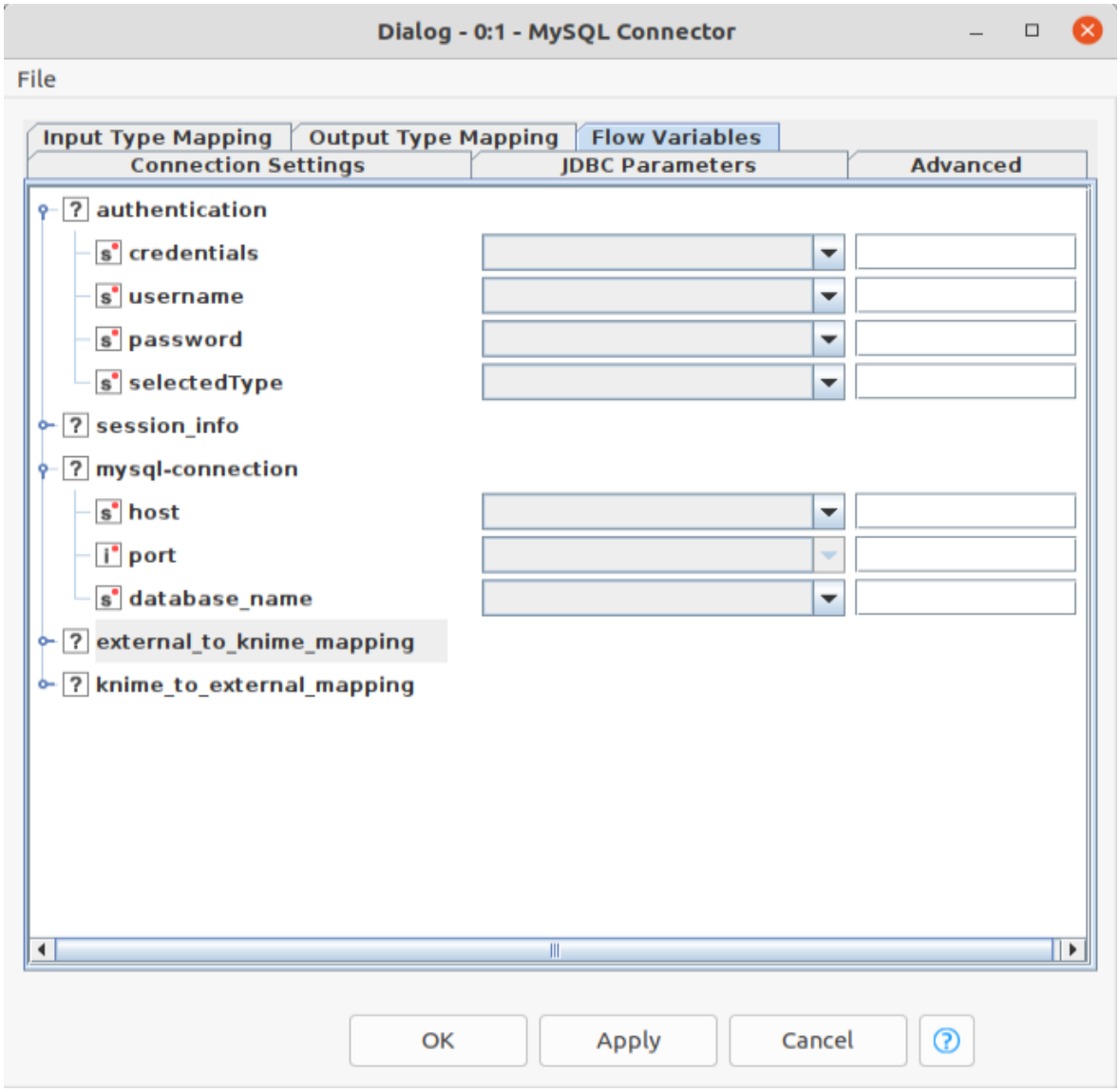
4-2 Połączenie Knime z bazą w mysql - Connection Settings



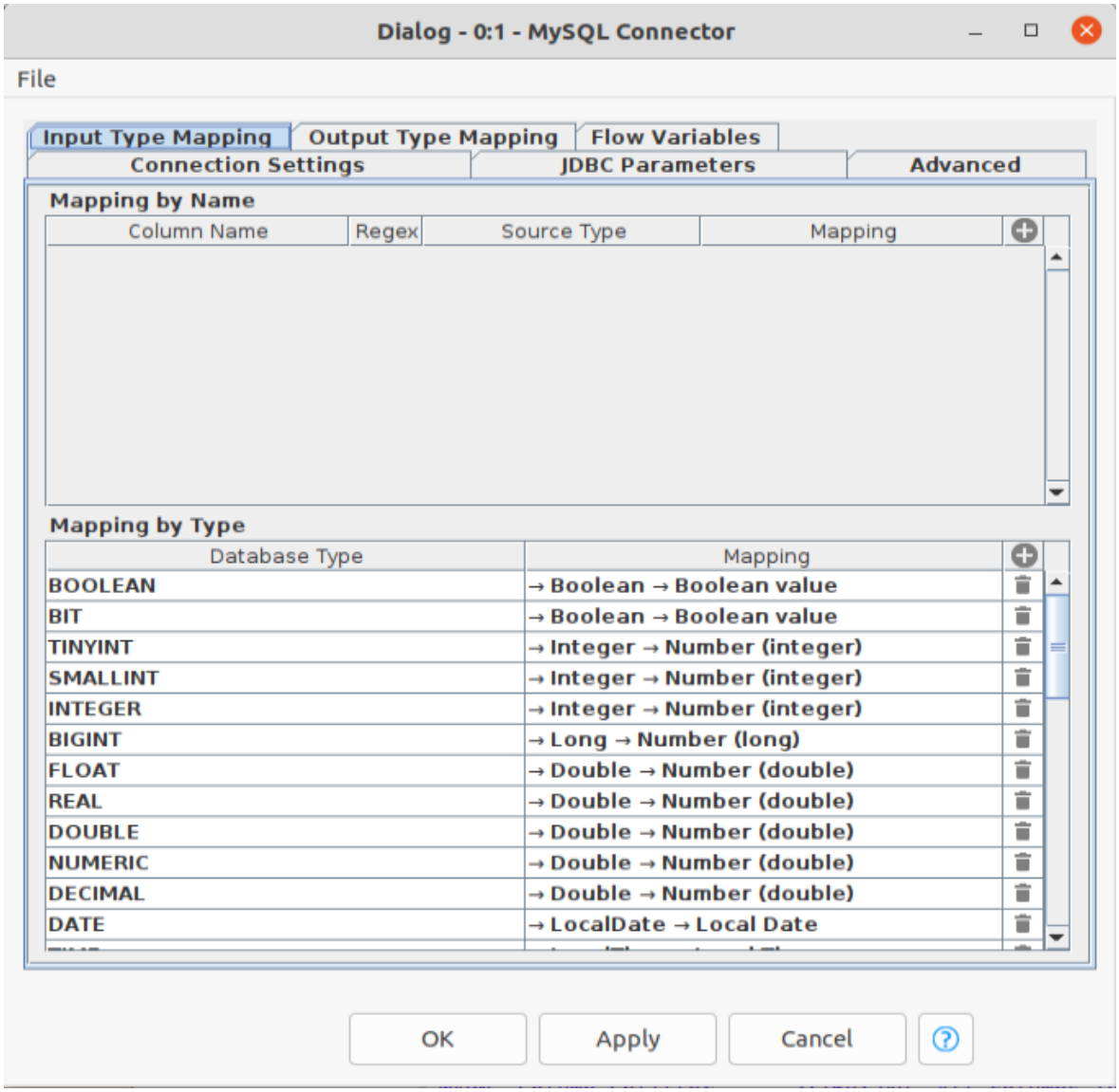
4-3 Połączenie Knime z bazą w mysql - Advanced



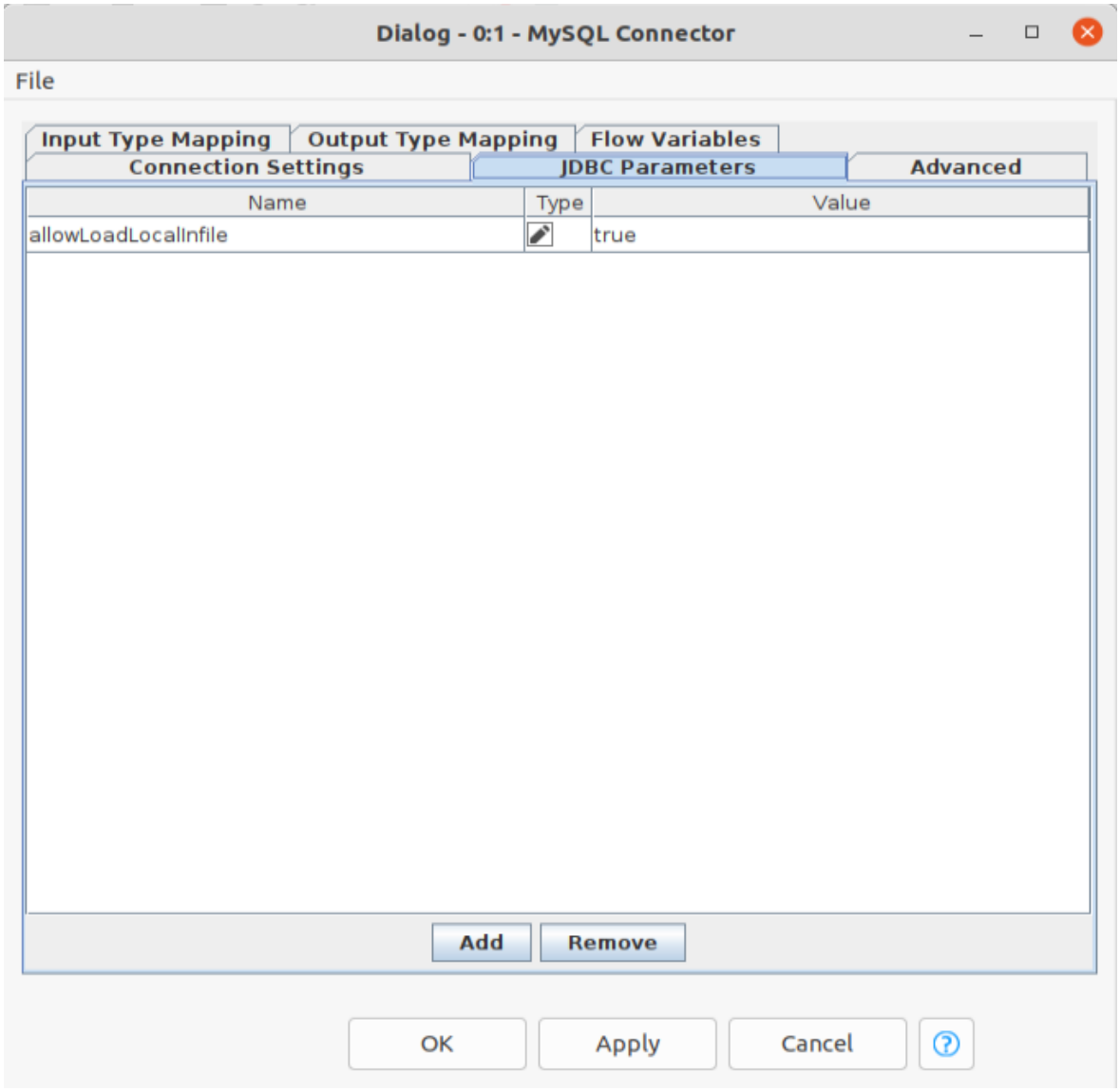
4-4 Połączenie Knime z bazą w mysql - Output Type Mapping



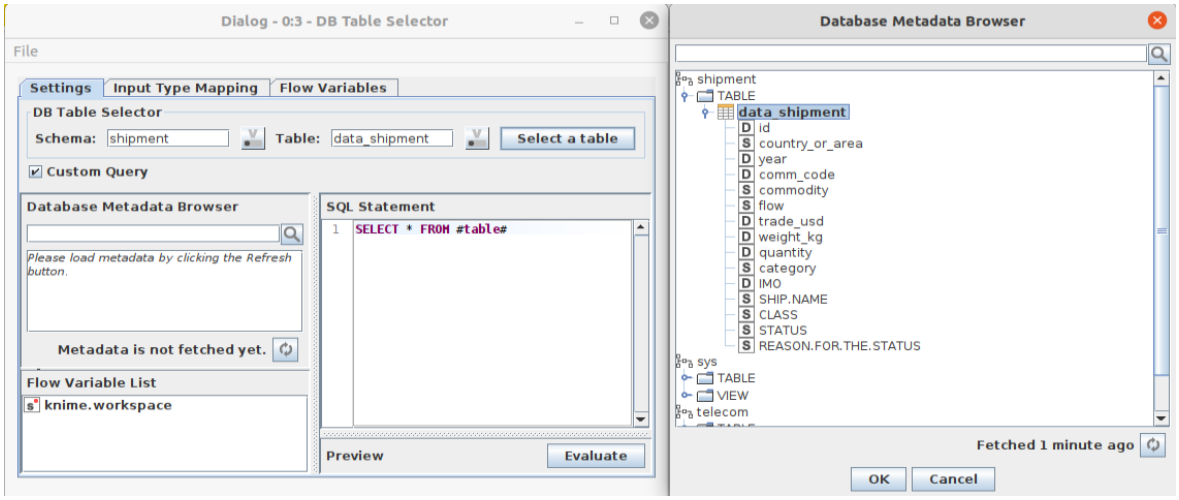
4-5 Połączenie Knime z bazą w mysql - Flow Variables



4-6 Połączenie Knime z bazą w mysql - Input Type Mapping

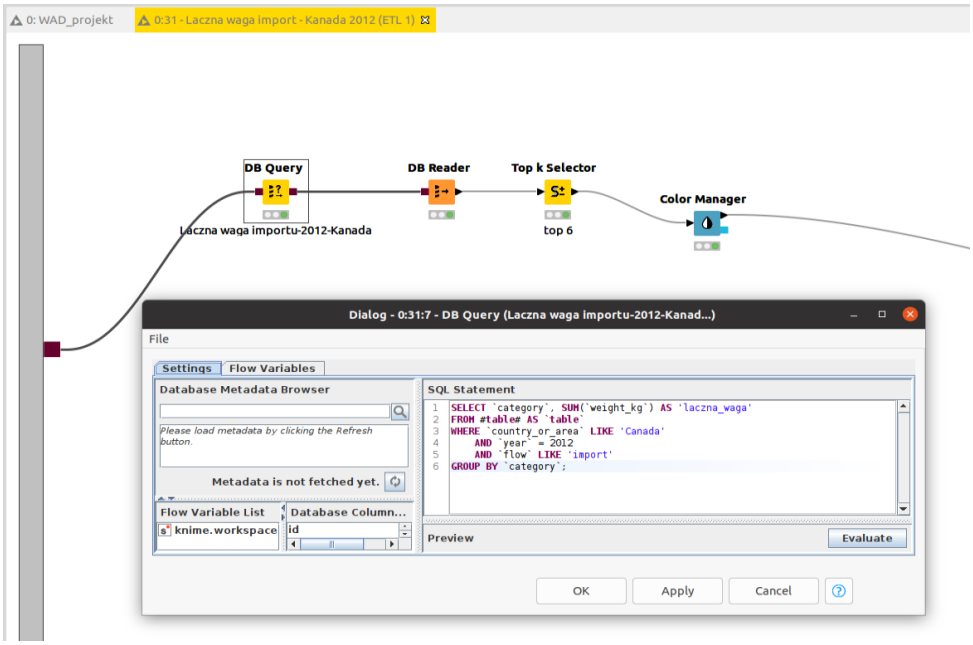


4-7 Połączenie Knime z bazą w mysql - JDBC Parameters

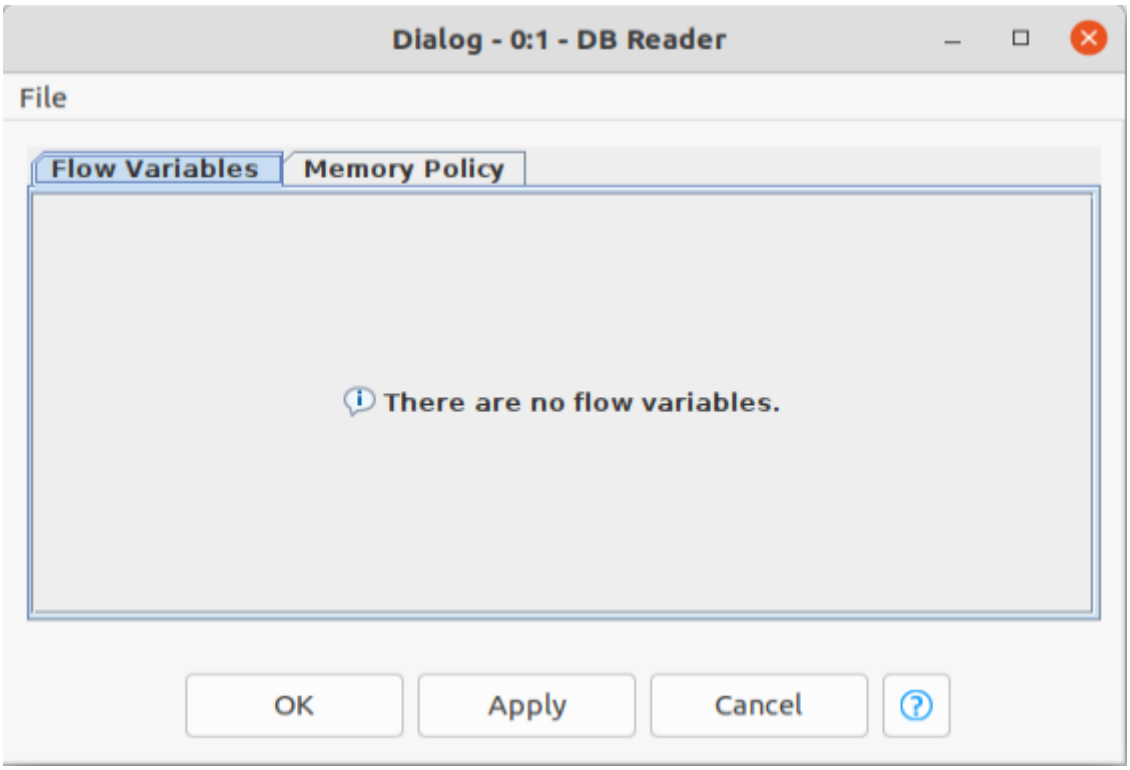


4-8 Table Selector – Settings

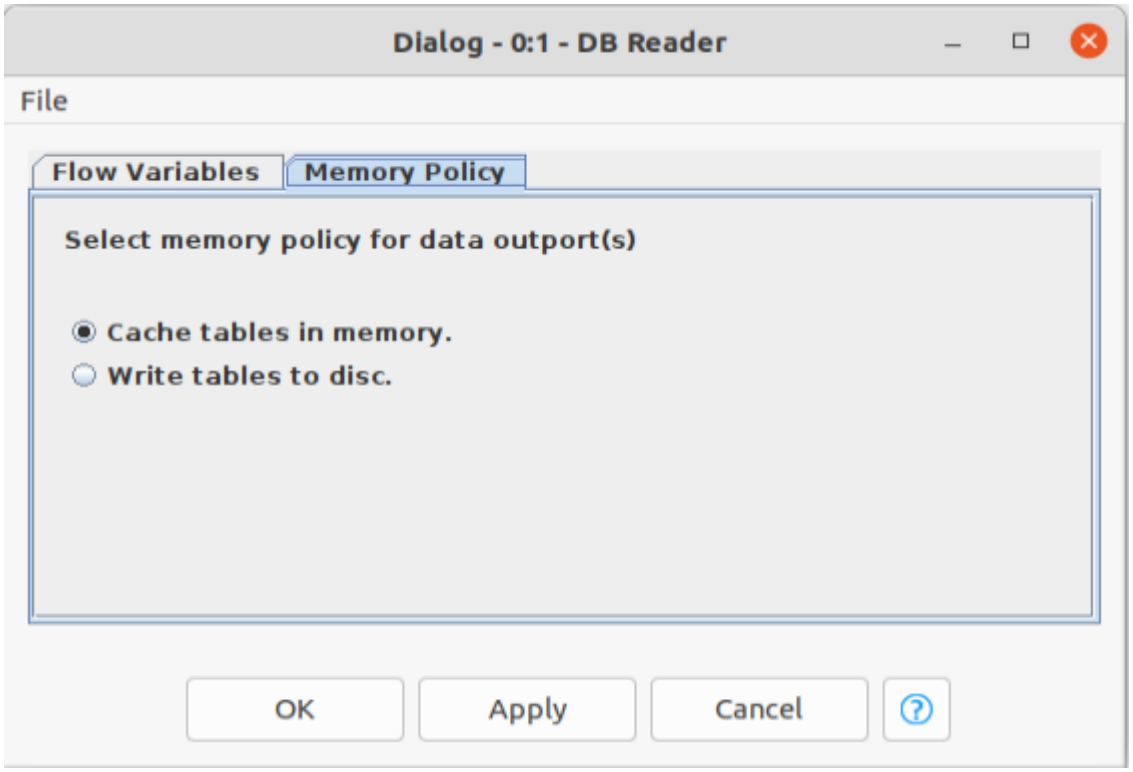
Poniżej znajduje się kilka z ETL, które wykonaliśmy przed tworzeniem wykresów. Pierwsza część dotyczy Kanady, druga natomiast porównania różnych krajów.



4-9 ETL z kanadyjskim importem w 2012 roku

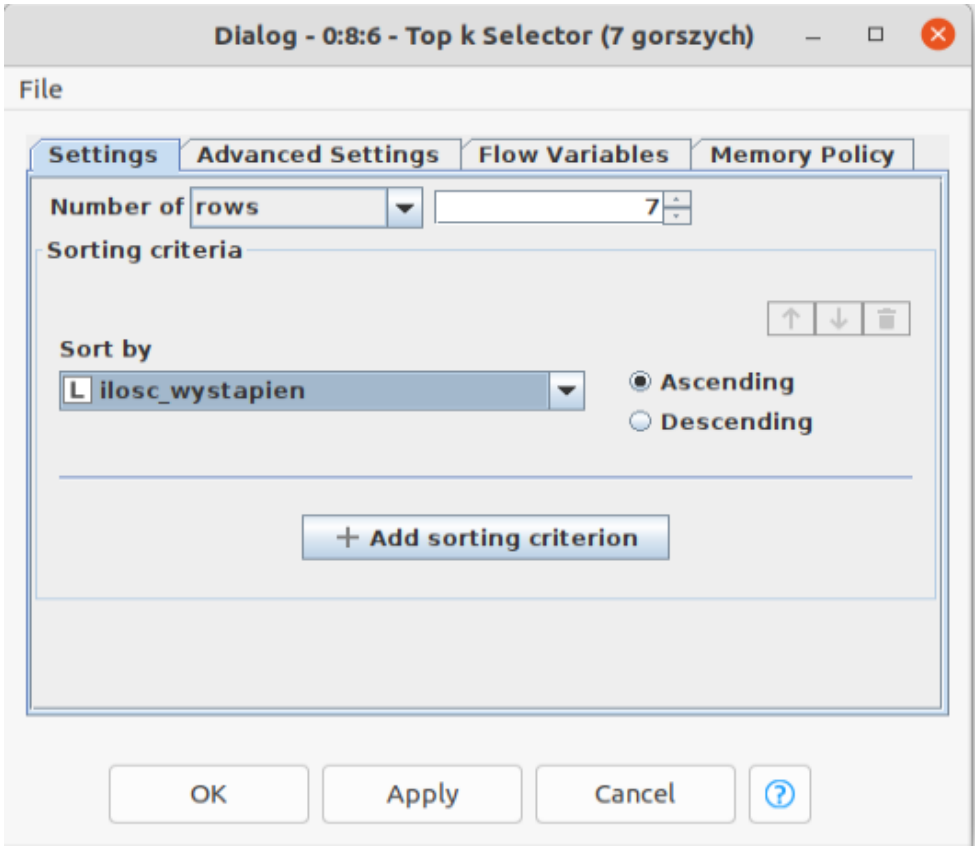


4-10 Okno edycji w węźle DB Reader (Flow Variables)

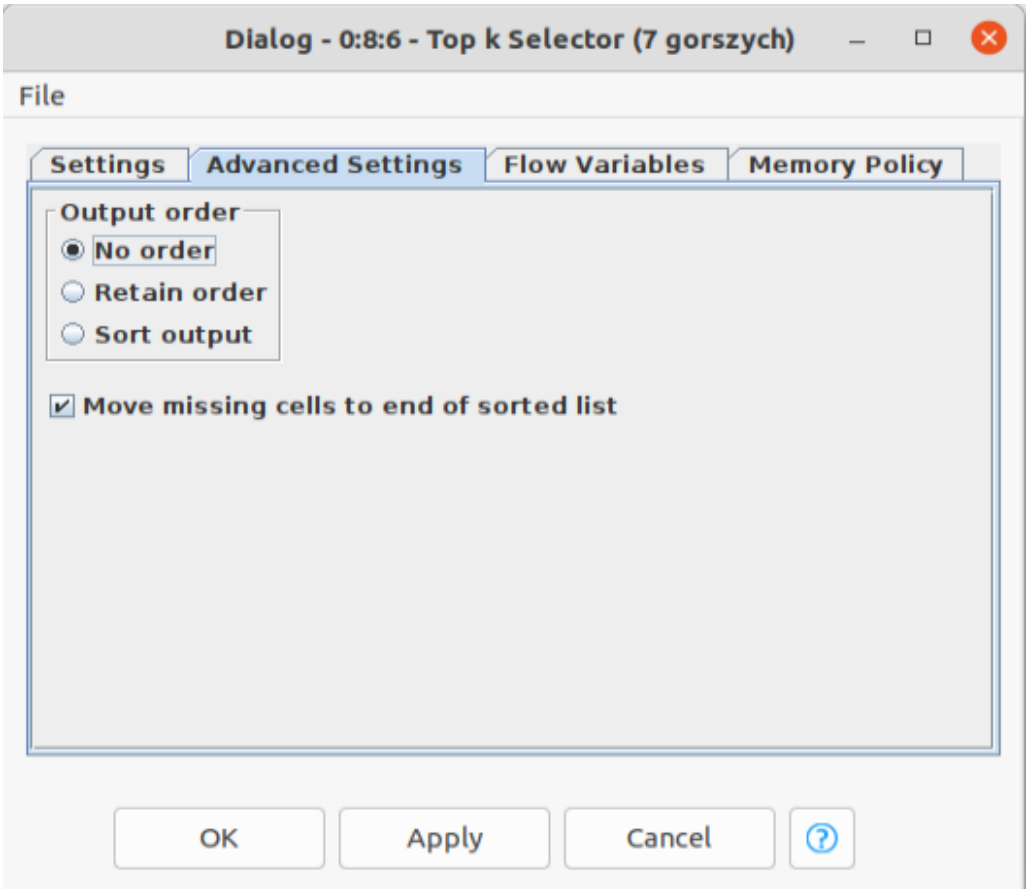


4-11 Okno edycji w węźle DB Reader (Memory Policy)

W węźle DB Reader nic nie ustawialiśmy służył do zmiany rodzaju przepływu, która była konieczna aby kontynuować pracę z innymi węzłami.

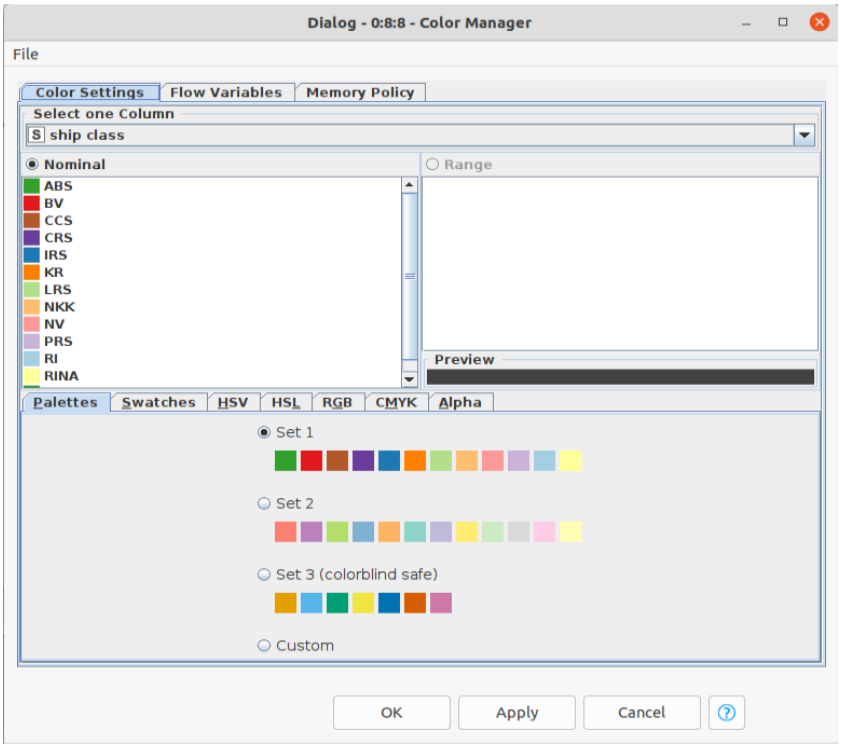


4-12 Top K Selector – Settings

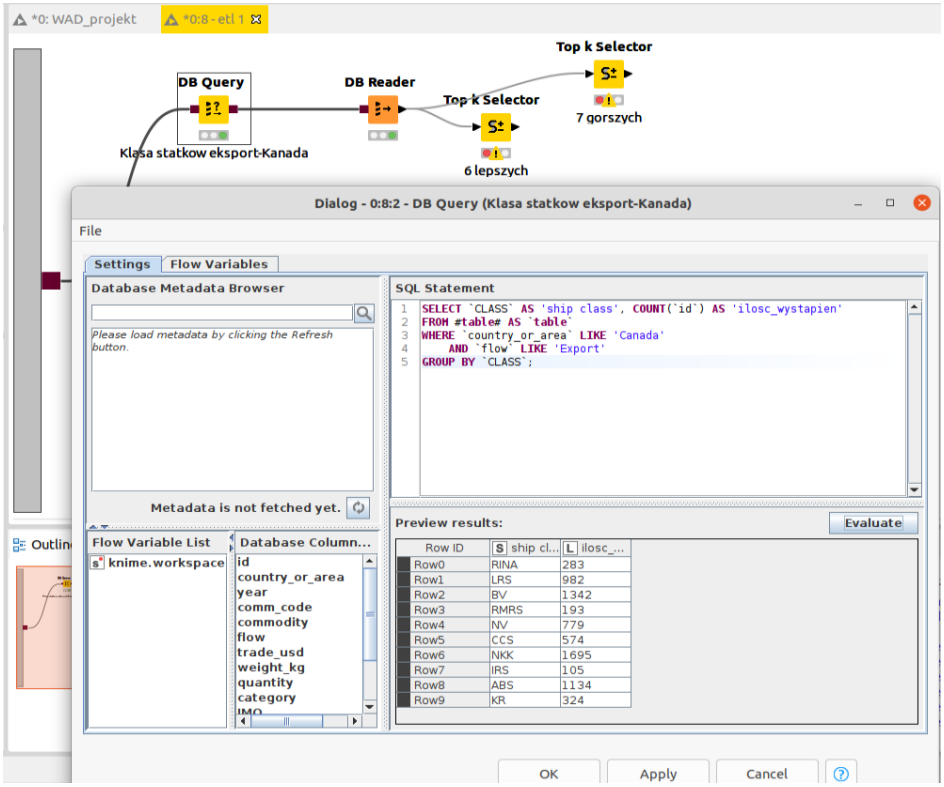


4-13 Top K Selector - Advanced Settings

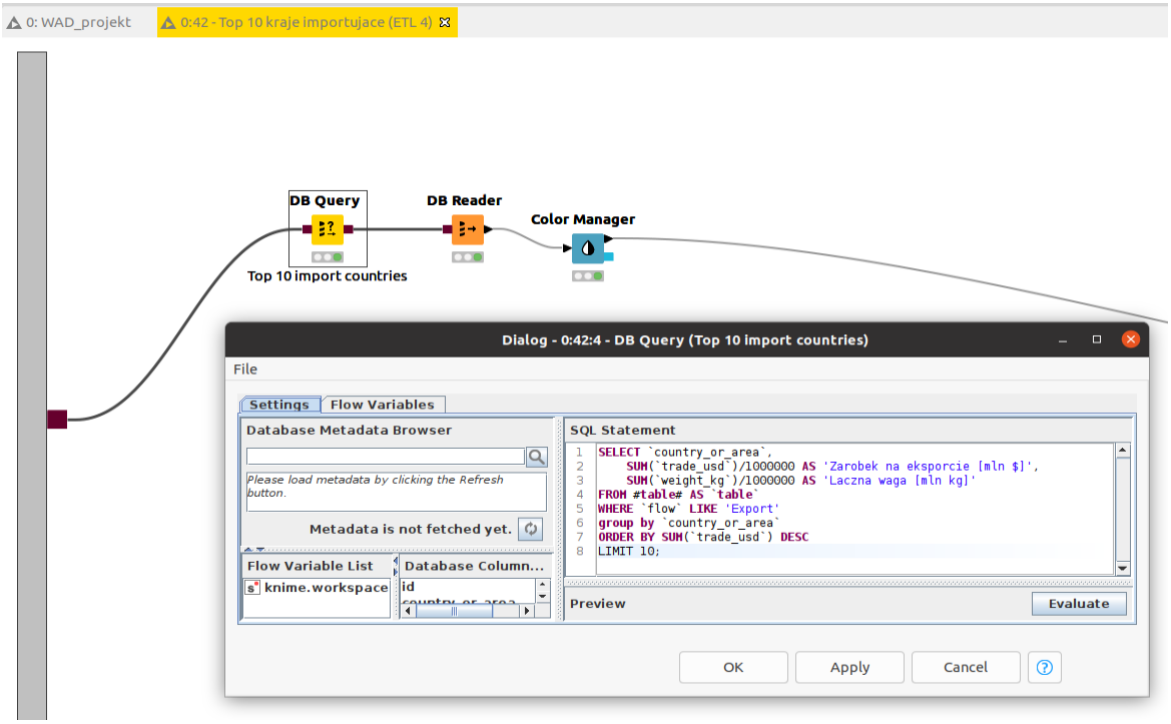
Top K Selector służy do ograniczenia liczby zwracanych wyników.



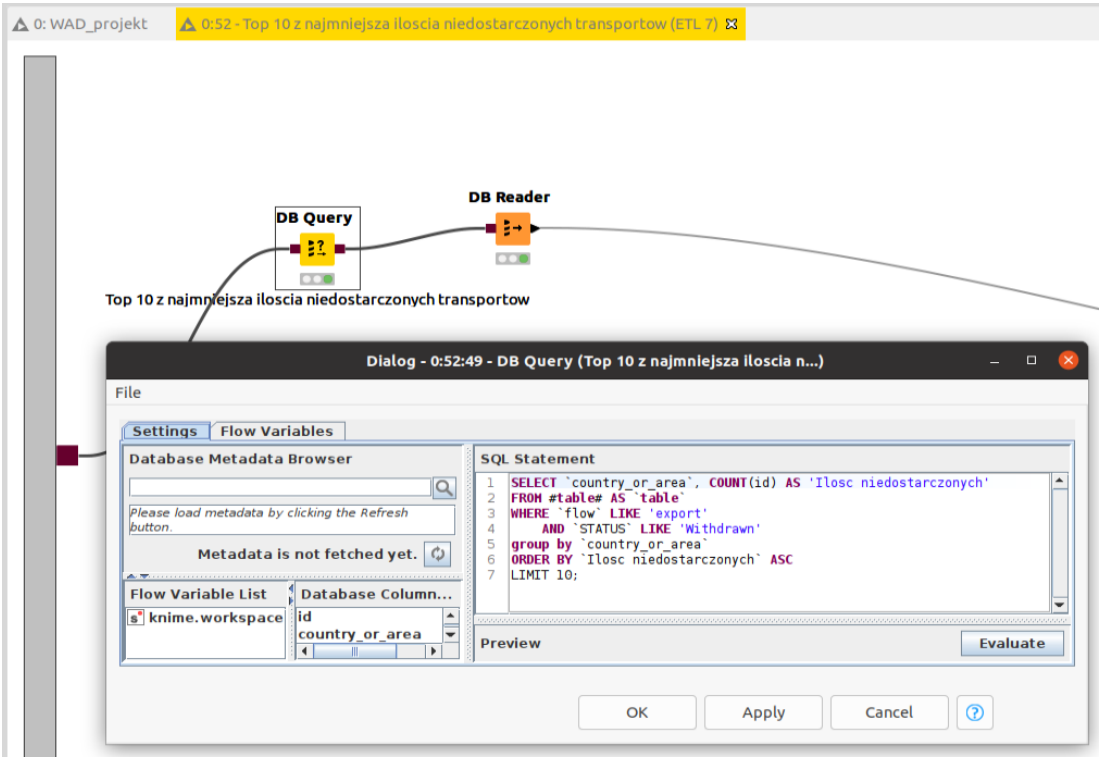
4-14 Color Manager



4-15 ETL z użyciem poszczególnych klas statków przez Kanadę w eksporcie



4-16 ETL z zarobkiem i wagą eksportu w poszczególnych krajach

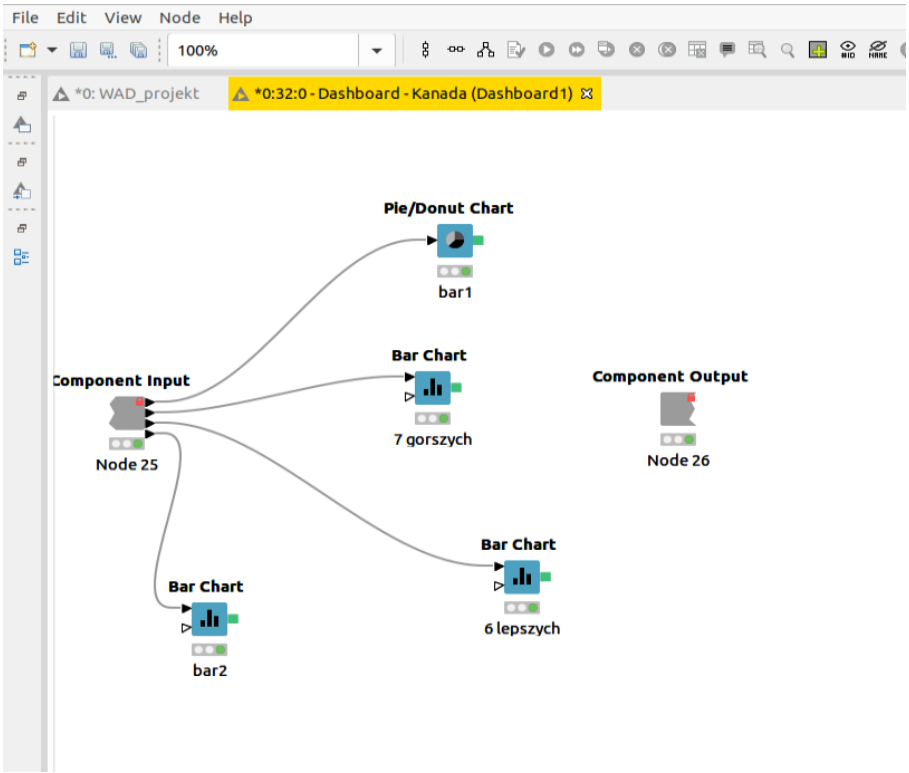


4-17 ETL z krajami, które mają najmniej niedostarczonych transportów

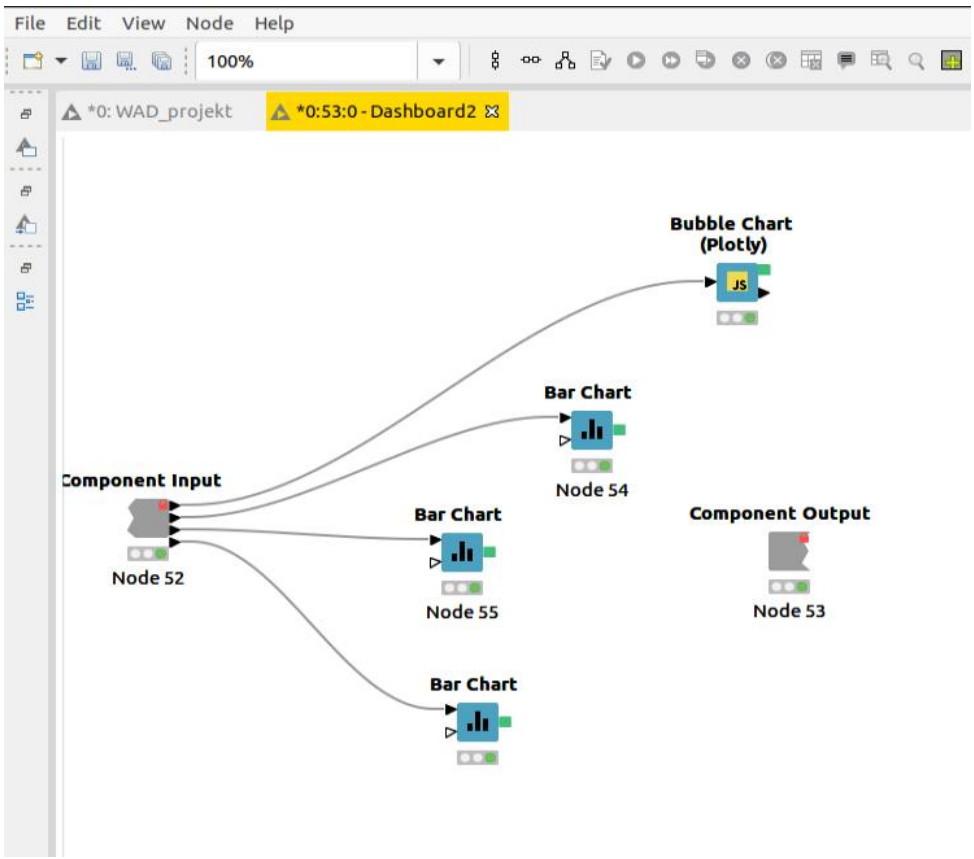
Jak można zauważyć, często korzystaliśmy po prostu z kwerend pisanych w SQL, ponieważ nasze dane były już na tyle przygotowane, że było to narzędzie wystarczające.

4.2. Dashboardy

Poniżej znajdują się dwa dashboardy, poprzedzone widokiem ich wykonania w Knime.



4-18 Budowa dashboardu dla Kanady



4-19 Budowa dashboardu z porównaniem krajów

4-20 Dashboard dla Kanady

4-21 Dashboard z porównaniem krajów

Dashboard dla Kanady odpowiada na pytania:

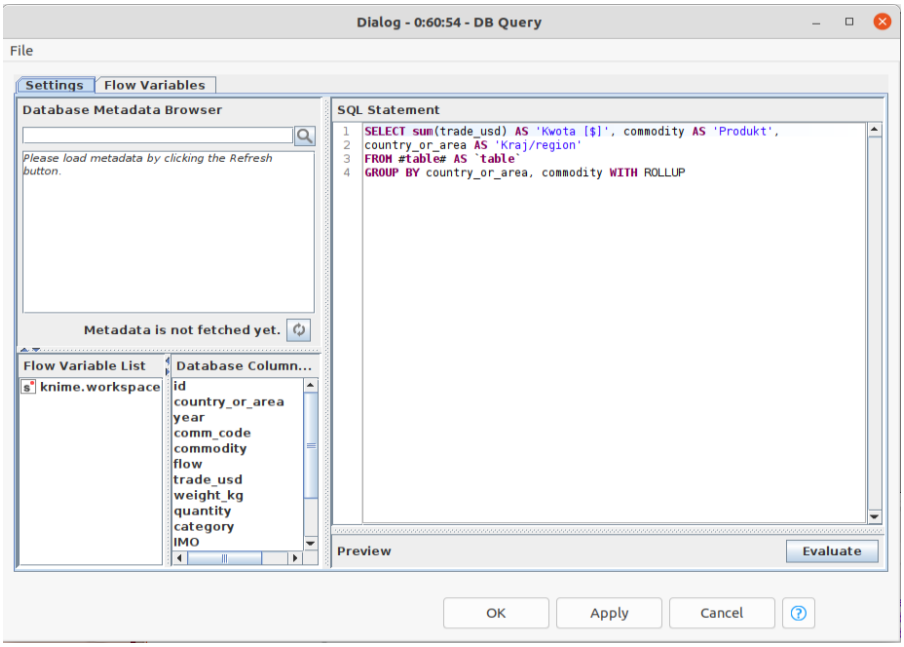
- Jakie klasy statków są najczęściej używane do eksportu w Kanadzie?
- Jakie klasy statków są najrzadziej używane do eksportu w Kanadzie?
- Na import jakich towarów Kanada wydaje najwięcej?
- Towarów których kategorii najwięcej importowała Kanada w 2012 roku?

Dashboard z porównaniem poszczególnych krajów odpowiada na pytania:

- Które kraje mają najmniejsze średnie wydatki na transport?
- Które kraje mają największe średnie wydatki na transport?
- Które kraje mają najmniej niedostarczonych transportów?
- Jak prezentuje się stosunek zarobku do wagi eksportu wśród krajów, które na eksporcie zarabiają najwięcej?

4.3. OLAP

Stworzyliśmy także trzy kostki OLAP., korzystając między innymi z klauzuli WITH ROLLUP. Kostki OLAP są to struktury danych, które pozwalają na szybką analizę danych. Przechowują one dane w sposób bardziej przypominający wielowymiarowe arkusze kalkulacyjne niż tradycyjną, relacyjną bazę danych



4-22 Kwerenda do stworzenia kostki z kwotą jaka obracana jest w handlu poszczególnymi produktami w danych krajach

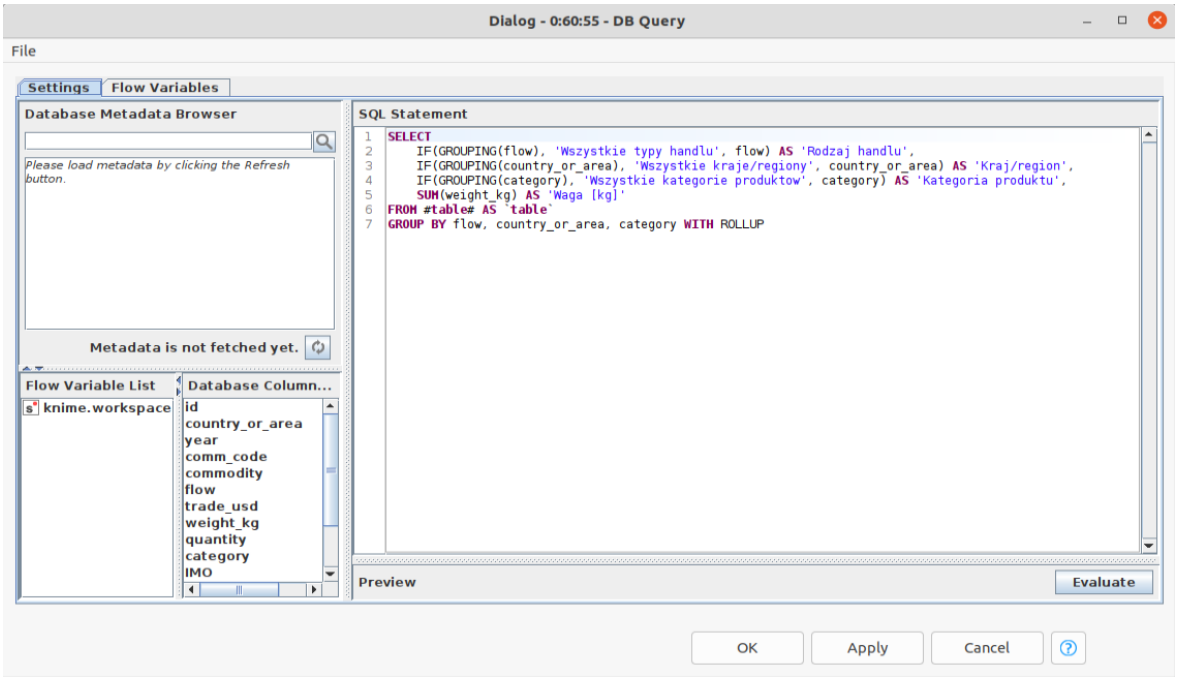
KNIME data table - 0:56 -

File Edit Hilite Navigation View

Table "database" - Rows: 247959 | Spec - Columns: 3 | Properties | Flow Variables

Row ID	D Kwota [$]	S Produkt	S Kraj/region
Row78	130,524,889	Refined sugar, in solid form, flavoured or coloured	Afghanistan
Row79	113,265,109	Retreaded tyres	Afghanistan
Row80	289,931	Rideable wheeled toys, dolls carriages	Afghanistan
Row81	3,868,724	Saffron	Afghanistan
Row82	3,731,294	Salt (sodium chloride) including solution, salt water	Afghanistan
Row83	1,217,158	Sandstone, merely cut into blocks etc	Afghanistan
Row84	25,312,153	Sections, U, iron or non-alloy steel, nfw hot-roll/drawn/ex...	Afghanistan
Row85	7,990,512	Seed, lucerne (alfalfa), for sowing	Afghanistan
Row86	50,233,366	Sesamum seeds	Afghanistan
Row87	64,796,247	Soaps nes	Afghanistan
Row88	60,363	Spices nes	Afghanistan
Row89	520,741,717	Stone setts, curbstones, flagstones (except slate)	Afghanistan
Row90	195,933,957	Tea, black (fermented or partly) in packages < 3 kg	Afghanistan
Row91	44,819,548	Tea, green (unfermented) in packages < 3 kg	Afghanistan
Row92	862,337	Telephone sets	Afghanistan
Row93	5,000	Telephonic or telegraphic switching apparatus	Afghanistan
Row94	21,887,951	Tomatoes, fresh or chilled	Afghanistan
Row95	2,149,002	Transmission apparatus for radio, telephone and TV	Afghanistan
Row96	186,827	Tungsten ores and concentrates	Afghanistan
Row97	358,500	Unit construction machines, metal work	Afghanistan
Row98	28,796,679	Urd,mung,black or green gram beans dried shelled	Afghanistan
Row99	85,858,029	Vegetable saps and extracts nes	Afghanistan
Row100	54,543,259	Vegetables, fresh or chilled nes	Afghanistan
Row101	4,031,480	Walnuts in shell, fresh or dried	Afghanistan
Row102	4,488,593	Walnuts, fresh or dried, shelled	Afghanistan
Row103	1,268,647,841	Wheat or meslin flour	Afghanistan
Row104	59,324,775	Womens, girls blouses, shirts, manmade fibre, not kni	Afghanistan
Row105	1,029,307	Wrist-watch, precious metal, battery, other	Afghanistan
Row106	1,963,539	Yogurt	Afghanistan
Row107	30,245,281,205	?	Afghanistan
Row108	?	"Other :-- Of a kind used in the leather or like industries	Albania
Row109	8,396	2,2'-oxydiethanol(diethylene glycol)	Albania
Row110	725,657	Abrasive powder or grain on a base of other material	Albania
Row111	1,413,089	Abrasive powder, grain on paper or paperboard support	Albania
Row112	288,676	Abrasive powderor grain on woven textile support	Albania
Row113	1,800,343	AC generators, of an output < 75 kVA	Albania
Row114	3,982,910	AC generators, of an output > 750 kVA	Albania
Row115	153,653	AC generators, of an output 375-750 kVA	Albania
Row116	107,218	AC generators, of an output 75-375 kVA	Albania
Row117	162,008	AC motors, multi-phase, of an output < 750 Watts	Albania
Row118	1,702,822	AC motors, multi-phase, of an output > 75 kW	Albania
Row119	1,077,330	AC motors, multi-phase, of an output 0.75-75 kW	Albania
Row120	537,710	AC motors, single-phase, nes	Albania
Row121	256,670	Acetic acid	Albania
Row122	200,862	Acetic acid esters nes	Albania
Row123	18,617	Acetic acid salts except cobalt and sodium	Albania
Row124	205,818	Acetone	Albania
Row125	158,010	Acid and mordant dyes and preparations based thereon	Albania
Row126	11,584,347	Acrylic & vinyl polymer based paint, varnish, in wate	Albania
Row127	5,517,339	Acrylic or vinyl polymer paint or varnish, non-aqueou	Albania

4-23 Widok kostki z podsumowaniem przepływu gotówki w handlu wszystkimi towarami..



4-24 Kwerenda do stworzenia kostki z informacjami na temat sumarycznej wagi ładunków w poszczególnych krajach zależnie od tego czy jest to import, eksport czy re-import

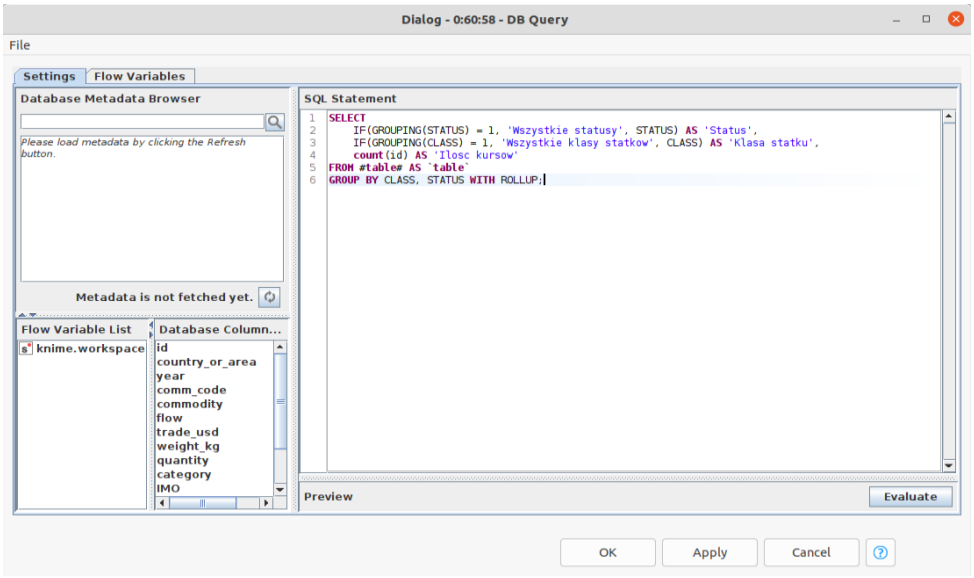
KNIME data table - 0:57 - DB Reader (Handel-Kraj-Kategoria-Waga)

File Edit Hilite Navigation View

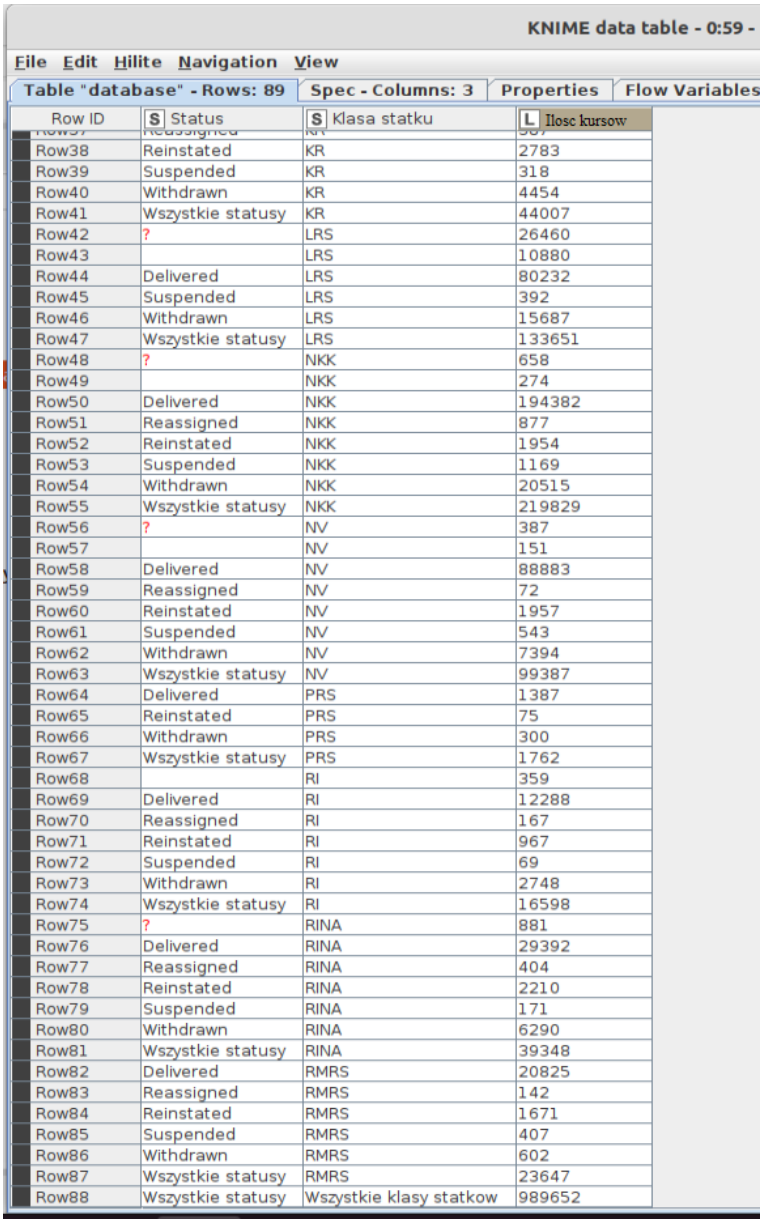
Table "database" - Rows: 26149 | Spec - Columns: 4 | Properties | Flow Variables

Row ID	Rodzaj handlu	Kraj/region	Kategoria produktu	Waga [kg]
Row26097	Re-Import	United Kingdom	10_cereals	385,369
Row26098	Re-Import	United Kingdom	13_lac_gums_resins_vegetable_saps_and_extracts_nes	13,132
Row26099	Re-Import	United Kingdom	14_vegetable_plaiting_materials_vegetable_products_nes	8,130
Row26100	Re-Import	United Kingdom	17_sugars_and_sugar_confectionery	355,161
Row26101	Re-Import	United Kingdom	18_cocoa_and_cocoa_preparations	368,263
Row26102	Re-Import	United Kingdom	19_cereal_flour_starch_milk_preparations_and_products	867,503
Row26103	Re-Import	United Kingdom	21_miscellaneous_edible_preparations	239,946
Row26104	Re-Import	United Kingdom	23_residues_wastes_of_food_industry_animal_fodder	27,916
Row26105	Re-Import	United Kingdom	24_tobacco_and_manufactured_tobacco_substitutes	6,337
Row26106	Re-Import	United Kingdom	26_ores_slag_and_ash	1,010
Row26107	Re-Import	United Kingdom	35_albuminoids_modified_starches_glues_enzymes	38,956
Row26108	Re-Import	United Kingdom	43_furskins_and_artificial_fur_manufactures_thereof	397
Row26109	Re-Import	United Kingdom	45_cork_and_articles_of_cork	16
Row26110	Re-Import	United Kingdom	46_manufactures_of_plaiting_material_basketwork_etc	19,800
Row26111	Re-Import	United Kingdom	47_pulp_of_wood_fibrous_cellulosic_material_waste_etc	17,016
Row26112	Re-Import	United Kingdom	50_silk	658
Row26113	Re-Import	United Kingdom	53_vegetable_textile_fibres_nes_paper_yarn_woven_fabri	1,453
Row26114	Re-Import	United Kingdom	60_knitted_or_crocheted_fabric	11,819
Row26115	Re-Import	United Kingdom	65_headgear_and_parts_thereof	101,784
Row26116	Re-Import	United Kingdom	66_umbrellas_walking_sticks_seat_sticks_whips_etc	125
Row26117	Re-Import	United Kingdom	67_bird_skin_feathers_artificial_flowers_human_hair	902
Row26118	Re-Import	United Kingdom	75_nickel_and_articles_thereof	149,422
Row26119	Re-Import	United Kingdom	78_lead_and_articles_thereof	11,969
Row26120	Re-Import	United Kingdom	79_zinc_and_articles_thereof	3,675
Row26121	Re-Import	United Kingdom	80_tin_and_articles_thereof	6,903
Row26122	Re-Import	United Kingdom	81_other_base_metals_cermets_articles_thereof	209,317
Row26123	Re-Import	United Kingdom	86_railway_tramway_locomotives_rolling_stock_equipmen	2,698,480
Row26124	Re-Import	United Kingdom	88_aircraft_spacecraft_and_parts_thereof	675,122
Row26125	Re-Import	United Kingdom	89_ships_boats_and_other_floating_structures	1,603,899
Row26126	Re-Import	United Kingdom	93_arms_and_ammunition_parts_and_accessories_thereof	3,727
Row26127	Re-Import	United Kingdom	97_works_of_art_collectors_pieces_and_antiques	1,155,837
Row26128	Re-Import	United Kingdom	99_commodities_not_specified_according_to_kind	?
Row26129	Re-Import	United Kingdom	all_commodities	?
Row26130	Re-Import	United Kingdom	Wszystkie kategorie produktow	11,891,678
Row26131	Re-Import	Uruguay	01_live_animals	4,600
Row26132	Re-Import	Uruguay	10_cereals	312,795
Row26133	Re-Import	Uruguay	19_cereal_flour_starch_milk_preparations_and_products	3,501
Row26134	Re-Import	Uruguay	23_residues_wastes_of_food_industry_animal_fodder	300
Row26135	Re-Import	Uruguay	43_furskins_and_artificial_fur_manufactures_thereof	155
Row26136	Re-Import	Uruguay	47_pulp_of_wood_fibrous_cellulosic_material_waste_etc	72,549,534
Row26137	Re-Import	Uruguay	60_knitted_or_crocheted_fabric	21,796
Row26138	Re-Import	Uruguay	86_railway_tramway_locomotives_rolling_stock_equipmen	1,189,344
Row26139	Re-Import	Uruguay	97_works_of_art_collectors_pieces_and_antiques	1,661
Row26140	Re-Import	Uruguay	all_commodities	?
Row26141	Re-Import	Uruguay	Wszystkie kategorie produktow	74,083,686
Row26142	Re-Import	Yemen	21_miscellaneous_edible_preparations	5,000
Row26143	Re-Import	Yemen	26_ores_slag_and_ash	65,000
Row26144	Re-Import	Yemen	93_arms_and_ammunition_parts_and_accessories_thereof	986
Row26145	Re-Import	Yemen	all_commodities	?
Row26146	Re-Import	Yemen	Wszystkie kategorie produktow	70,986
Row26147	Re-Import	Wszystkie kraje/regiony	Wszystkie kategorie produktow	6,807,171,2...
Row26148	Wszystkie typy handlu	Wszystkie kraje/regiony	Wszystkie kategorie produktow	45,622,738,...

4-25 Kostka na temat handlu w poszczególnych krajach

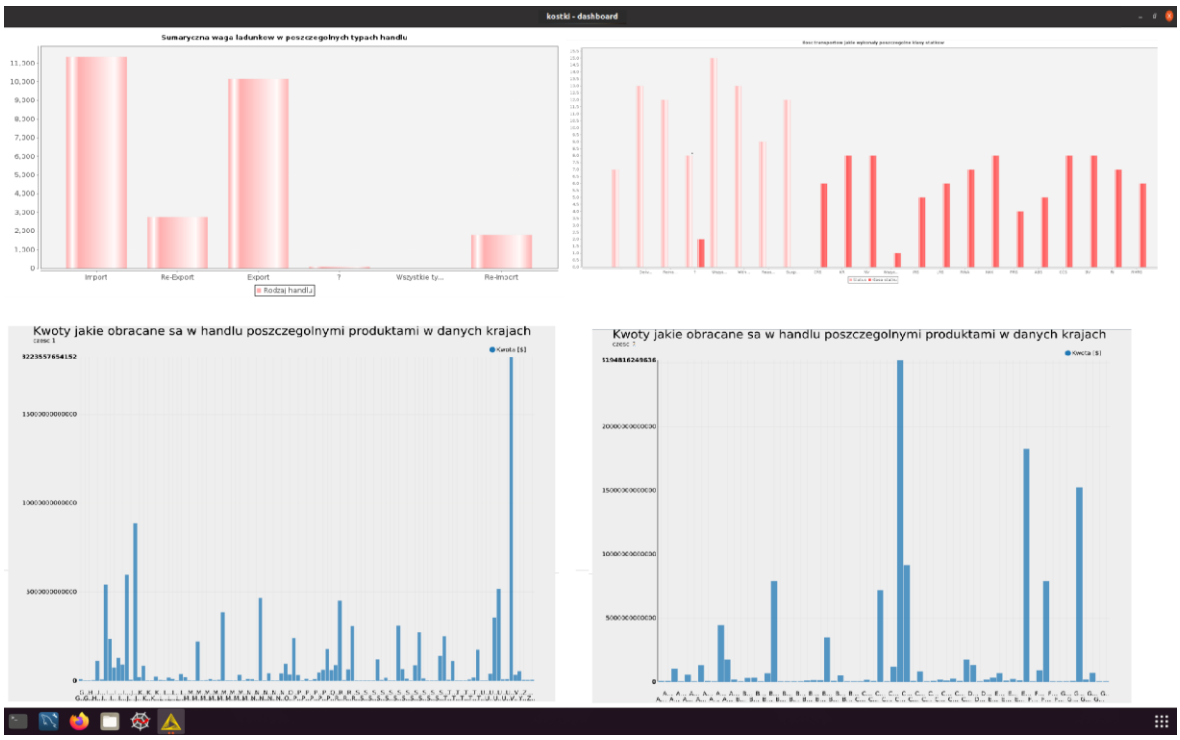


4-26 Kwerenda do stworzenia kostki z informacjami na temat transportów w kontekście klas statków



Row ID	Status	Klasa statku	Ilosc kursow
Row38	Reinstated	KR	2783
Row39	Suspended	KR	318
Row40	Withdrawn	KR	4454
Row41	Wszystkie statusy	KR	44007
Row42	?	LRS	26460
Row43		LRS	10880
Row44	Delivered	LRS	80232
Row45	Suspended	LRS	392
Row46	Withdrawn	LRS	15687
Row47	Wszystkie statusy	LRS	133651
Row48	?	NKK	658
Row49		NKK	274
Row50	Delivered	NKK	194382
Row51	Reassigned	NKK	877
Row52	Reinstated	NKK	1954
Row53	Suspended	NKK	1169
Row54	Withdrawn	NKK	20515
Row55	Wszystkie statusy	NKK	219829
Row56	?	NV	387
Row57		NV	151
Row58	Delivered	NV	88883
Row59	Reassigned	NV	72
Row60	Reinstated	NV	1957
Row61	Suspended	NV	543
Row62	Withdrawn	NV	7394
Row63	Wszystkie statusy	NV	99387
Row64	Delivered	PRS	1387
Row65	Reinstated	PRS	75
Row66	Withdrawn	PRS	300
Row67	Wszystkie statusy	PRS	1762
Row68		RI	359
Row69	Delivered	RI	12288
Row70	Reassigned	RI	167
Row71	Reinstated	RI	967
Row72	Suspended	RI	69
Row73	Withdrawn	RI	2748
Row74	Wszystkie statusy	RI	16598
Row75	?	RINA	881
Row76	Delivered	RINA	29392
Row77	Reassigned	RINA	404
Row78	Reinstated	RINA	2210
Row79	Suspended	RINA	171
Row80	Withdrawn	RINA	6290
Row81	Wszystkie statusy	RINA	39348
Row82	Delivered	RMRS	20825
Row83	Reassigned	RMRS	142
Row84	Reinstated	RMRS	1671
Row85	Suspended	RMRS	407
Row86	Withdrawn	RMRS	602
Row87	Wszystkie statusy	RMRS	23647
Row88	Wszystkie statusy	Wszystkie klasy statkow	989652

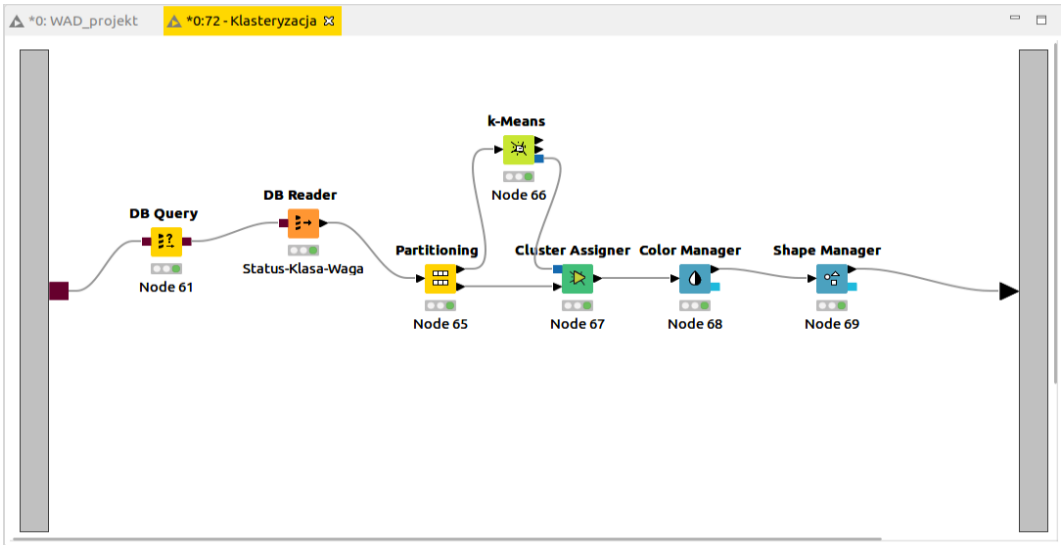
4-27 Kostka z informacjami na temat transportów i wykorzystywanych statków



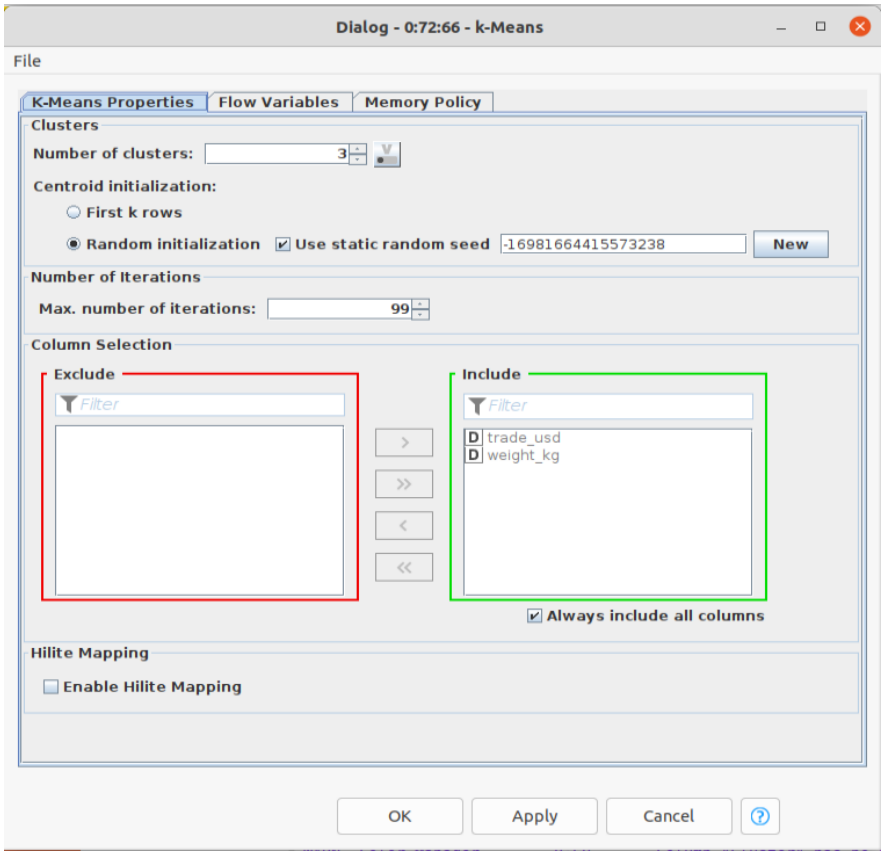
4-28 Dashboard zrobiony na podstawie kostek OLAP

4.4. Data mining

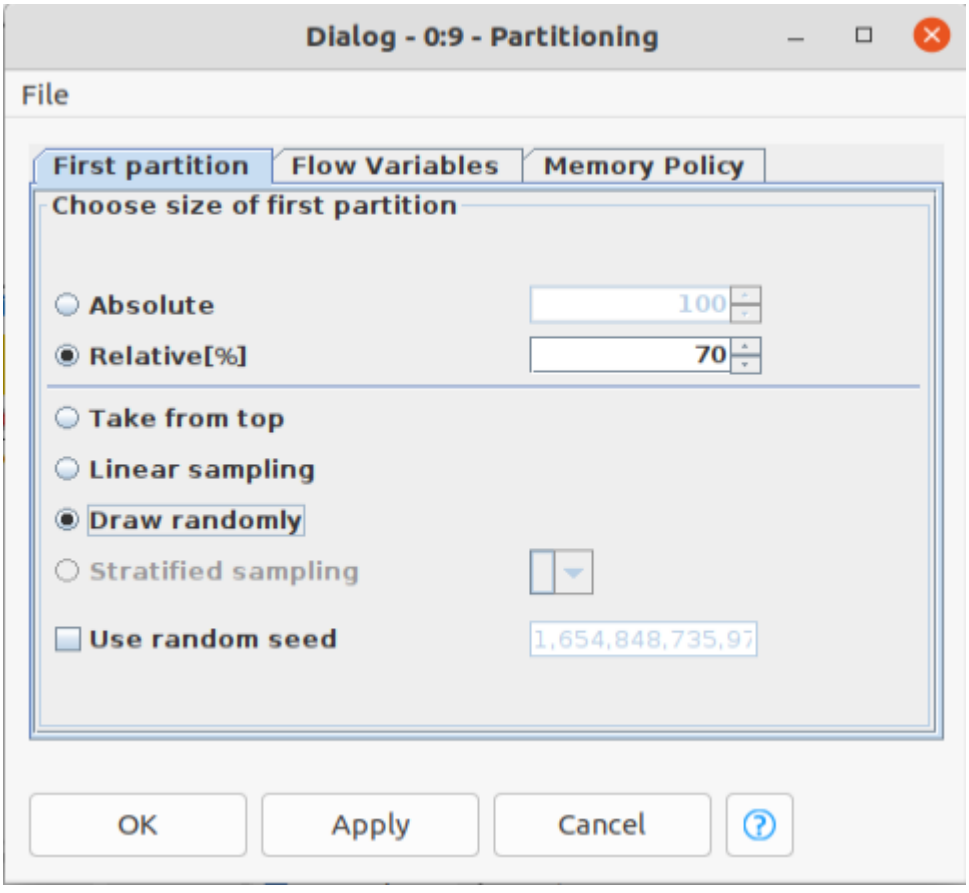
Wykorzystaliśmy także klasteryzację k-Means jako jeden z elementów data mining.



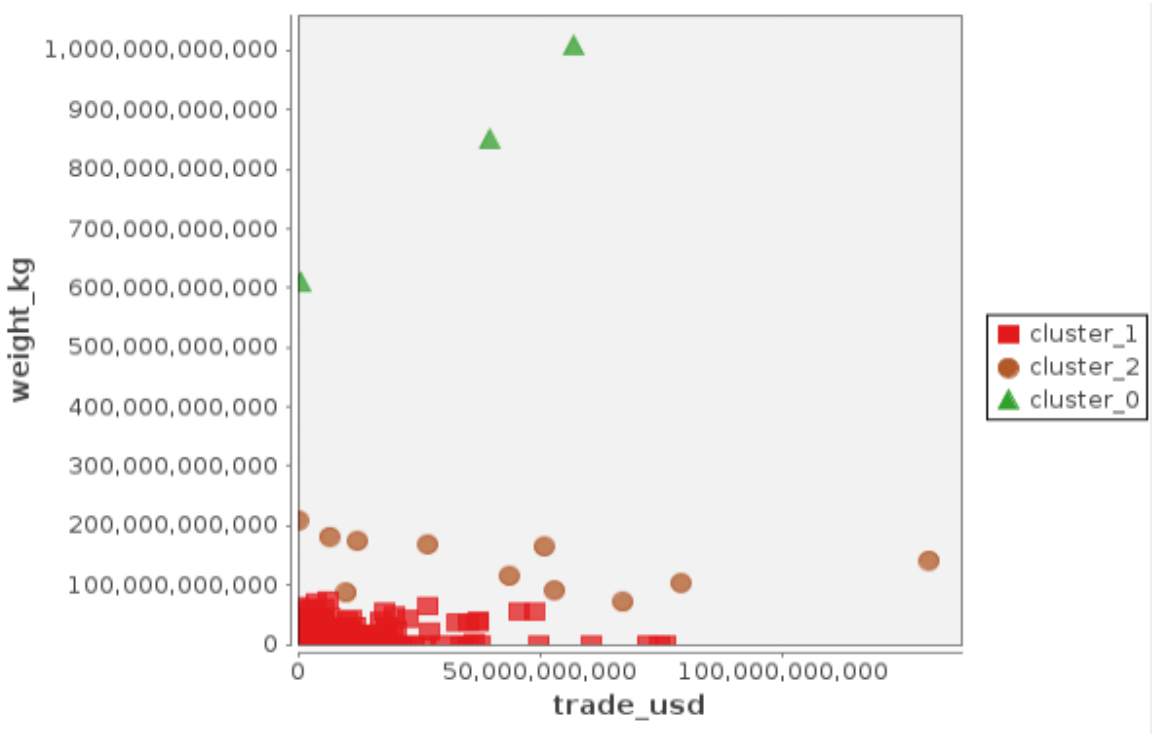
4-29 Budowa klasteryzacji



4-30 Ustawienie parametrów dla algorytmu k-Means

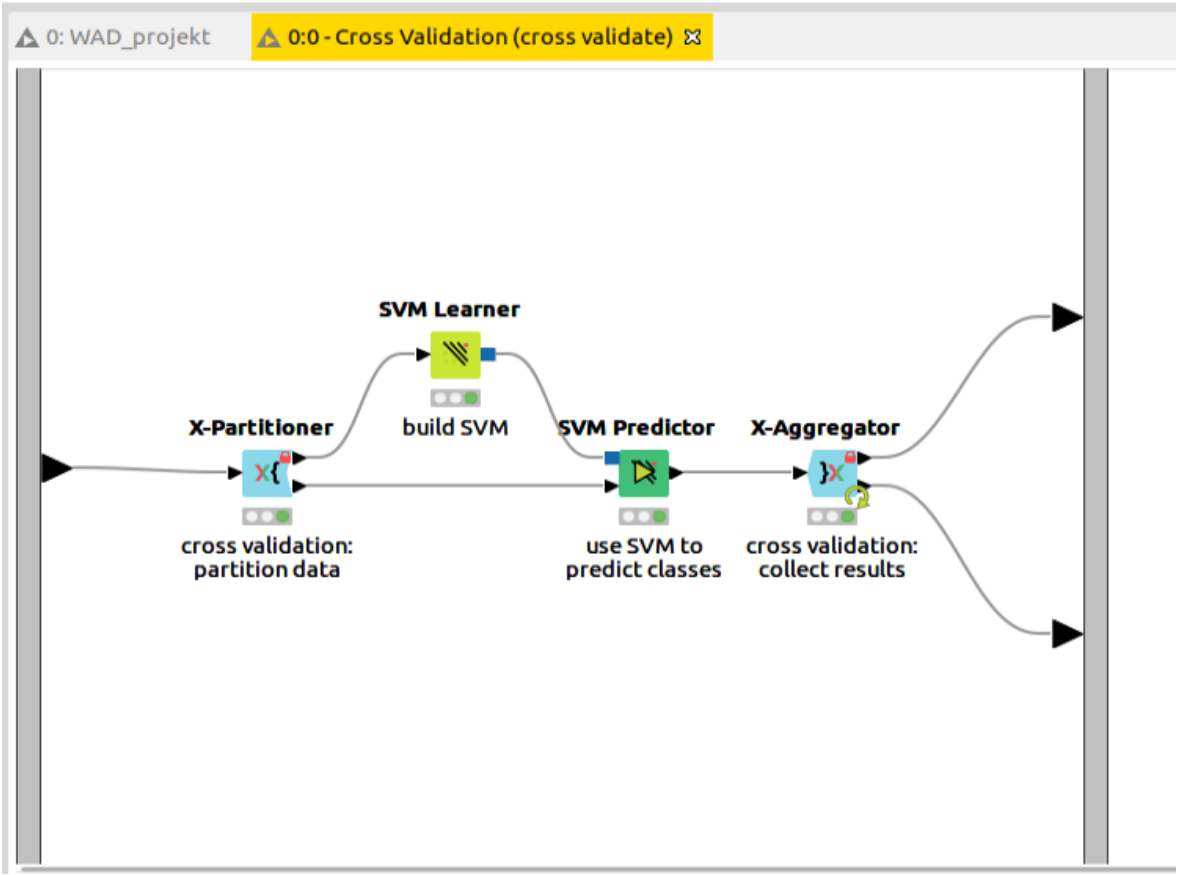


4-31 Wnętrze węzła Partitioning

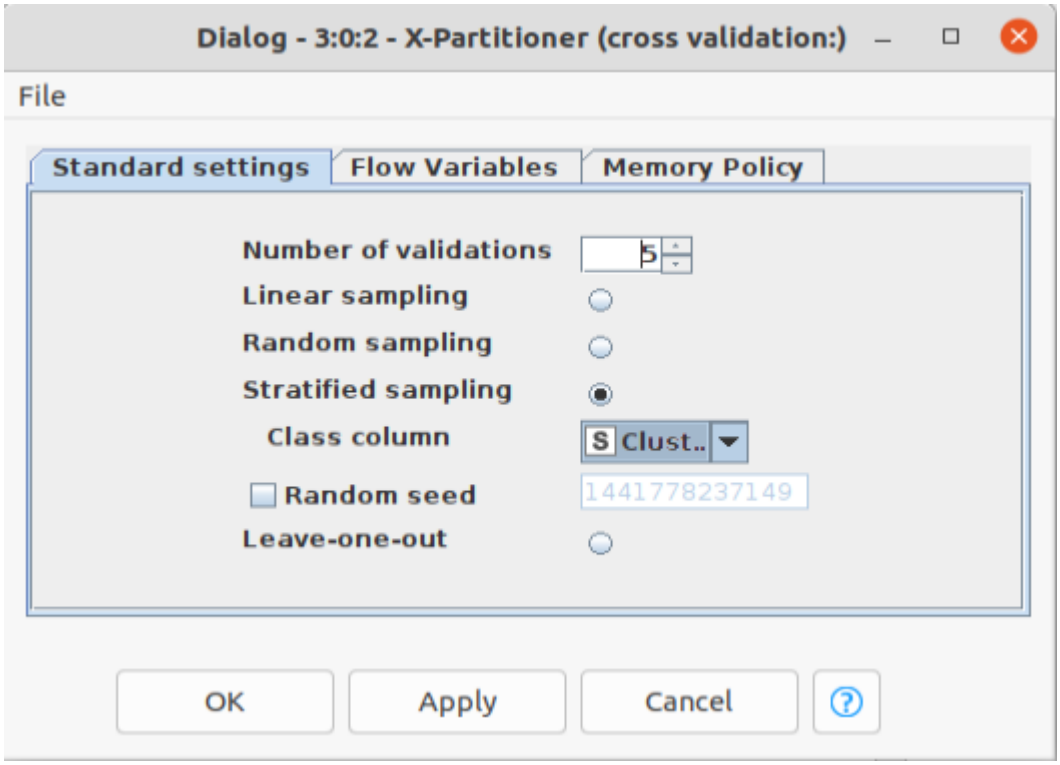


4-32 Wykres przedstawiający wyniki klasteryzacji

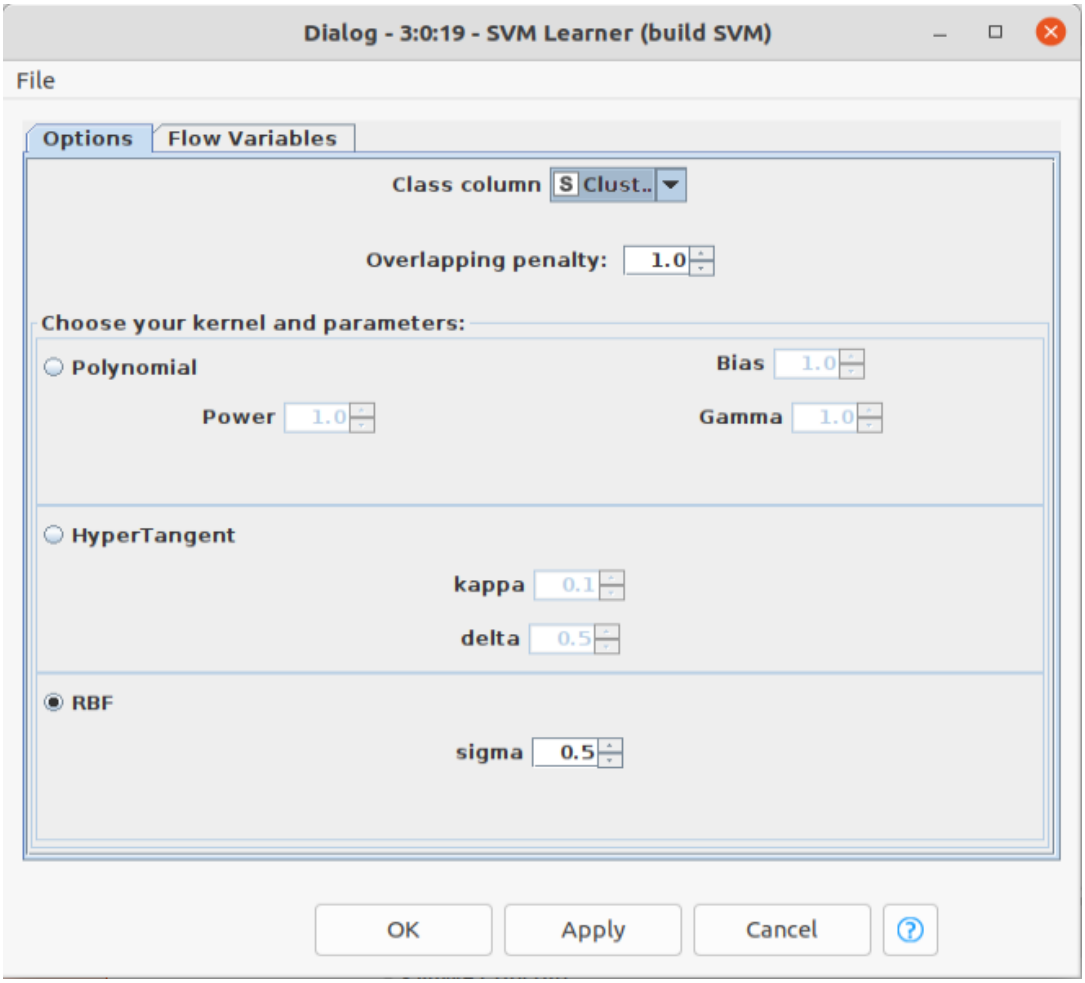
Większość danych została sklasyfikowana to pierwszego klastra, oznaczonego poprzez czerwone kwadraty. Na danych po klasteryzacji wykorzystaliśmy algorytm wektorów spierających SVM.



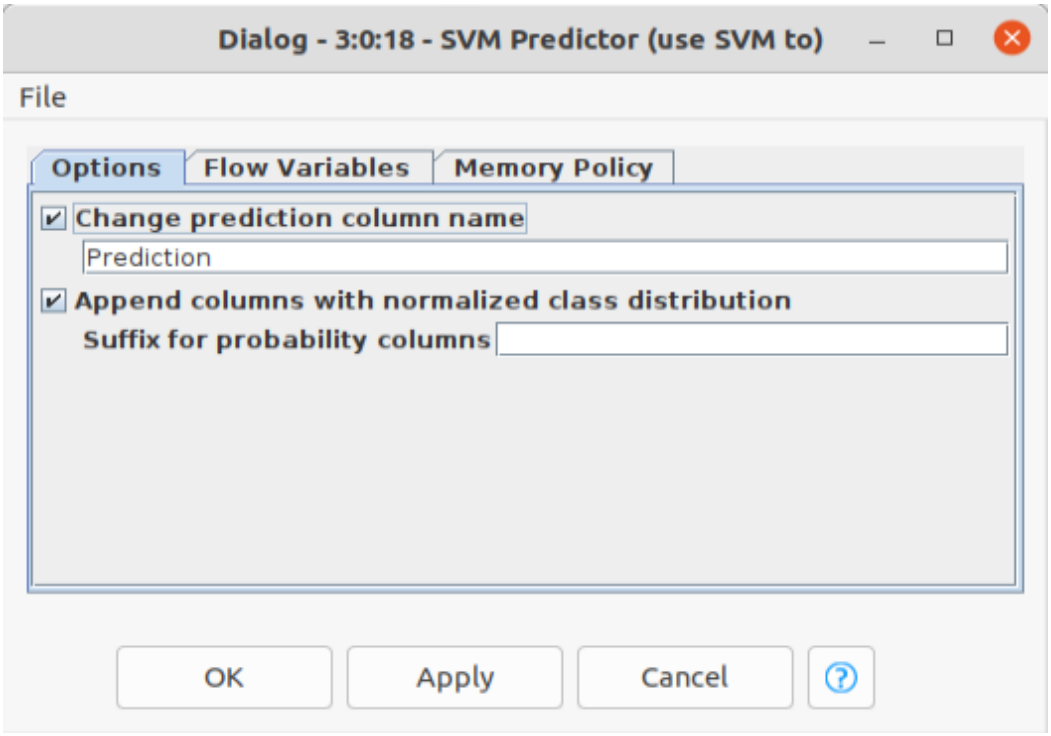
4-33 Budowa modelu SVM



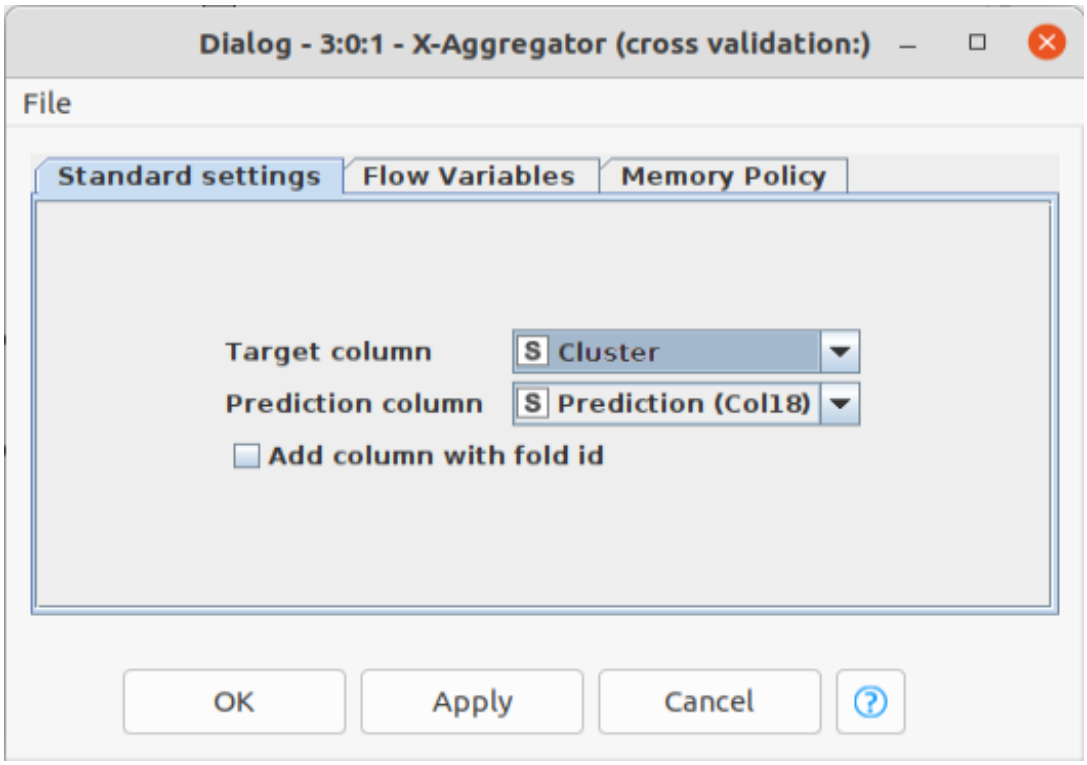
4-34 Wnętrze węzła X-Partitioner służącego do podziału danych walidacji krzyżowej



4-35 Wnętrze węzła SVM Learner



4-36 Wnętrze węzła SVM Predictor



4-37 Wnętrze węzła X-Aggregator służącego do zebrania wyników w walidacji krzyżowej

Confusion matrix - 0:16 - Scorer (deprecated) (score results)

File Edit Hilite Navigation View

Table "spec_name" - Rows: 3 | Spec - Columns: 3 | Properties | Flow Variables

Row ID	cluster_1	cluster_2	cluster_0
cluster_1	288428	11	3
cluster_2	0	0	0
cluster_0	0	0	0

4-38 Przyporządkowanie

Error rates - 0:0:1 - X-Aggregator (cross validation:)

File Edit Hilite Navigation View

Table "default" - Rows: 6 | Spec - Columns: 3 | Properties | Flow Variables

Row ID	Error in %	Size of Test Set	Error Count
fold 0	0.005	144221	7
fold 1	0.005	144221	7
fold 2	0.005	144221	7
fold 3	0.005	144221	7
fold 4	0.005	144221	7
fold 5	0.005	144221	7

4-39 Zliczanie błędów

Prediction table - 0:0:1 - X-Aggregator (cross validation:)

File Edit Hilite Navigation View

Table "default" - Rows: 288442 | Spec - Columns: 7 | Properties | Flow Variables

Row ID	trade_...	weight...	Cluster	P (Cluster=cluster_1)	P (Cluster=cluster_2)	P (Cluster=cluster_0)	Prediction
Row3	1,400,027	4,081,573	cluster_1	0.576	0.212	0.212	cluster_1
Row8	1,475,69...	3,893,67...	cluster_1	0.576	0.212	0.212	cluster_1
Row13	53,799,291	167,081,...	cluster_1	0.576	0.212	0.212	cluster_1
Row24	199,545,...	681,484,...	cluster_1	0.576	0.212	0.212	cluster_1
Row31	247,836,...	264,193,...	cluster_1	0.576	0.212	0.212	cluster_1
Row41	9,109	5,545	cluster_1	0.576	0.212	0.212	cluster_1
Row42	91,708,090	120,977,...	cluster_1	0.576	0.212	0.212	cluster_1
Row46	12,626	2,148	cluster_1	0.576	0.212	0.212	cluster_1
Row49	9,512	10,889	cluster_1	0.576	0.212	0.212	cluster_1
Row54	10,432	1,071	cluster_1	0.576	0.212	0.212	cluster_1
Row60	28,969,085	2,538,792	cluster_1	0.576	0.212	0.212	cluster_1
Row61	257,139,...	81,383,205	cluster_1	0.576	0.212	0.212	cluster_1
Row65	45,658	27,093	cluster_1	0.576	0.212	0.212	cluster_1
Row67	60,791,499	93,830,277	cluster_1	0.576	0.212	0.212	cluster_1
Row78	374,819,...	996,238,...	cluster_1	0.576	0.212	0.212	cluster_1
Row83	28,376,203	82,838,543	cluster_1	0.576	0.212	0.212	cluster_1
Row85	5,312,705	368,786	cluster_1	0.576	0.212	0.212	cluster_1
Row97	23,312,760	86,602,391	cluster_1	0.576	0.212	0.212	cluster_1
Row115	30,993	85,649	cluster_1	0.576	0.212	0.212	cluster_1
Row124	41,441	90,500	cluster_1	0.576	0.212	0.212	cluster_1
Row130	1,393	600	cluster_1	0.576	0.212	0.212	cluster_1
Row131	1,393	600	cluster_1	0.576	0.212	0.212	cluster_1
Row136	16,524,095	23,719,022	cluster_1	0.576	0.212	0.212	cluster_1
Row152	74,631,150	96,200,477	cluster_1	0.576	0.212	0.212	cluster_1
Row153	3	2	cluster_1	0.576	0.212	0.212	cluster_1
Row155	5,414	20,000	cluster_1	0.576	0.212	0.212	cluster_1
Row158	6,933,959	9,742,870	cluster_1	0.576	0.212	0.212	cluster_1
Row175	1,509,175	4,194,071	cluster_1	0.576	0.212	0.212	cluster_1
Row207	168,786	2,938,980	cluster_1	0.576	0.212	0.212	cluster_1
Row212	249	119	cluster_1	0.576	0.212	0.212	cluster_1
Row224	17,835	54,442	cluster_1	0.576	0.212	0.212	cluster_1
Row237	3,313,410	9,634,806	cluster_1	0.576	0.212	0.212	cluster_1
Row238	27,064,207	103,210,...	cluster_1	0.576	0.212	0.212	cluster_1
Row254	4,970,067	12,106,573	cluster_1	0.576	0.212	0.212	cluster_1
Row258	41,641	56,931	cluster_1	0.576	0.212	0.212	cluster_1
Row263	6,258,771	5,609,607	cluster_1	0.576	0.212	0.212	cluster_1
Row266	2,282	600	cluster_1	0.576	0.212	0.212	cluster_1
Row268	233,172	267,573	cluster_1	0.576	0.212	0.212	cluster_1
Row273	290,488,...	1,077,61...	cluster_1	0.576	0.212	0.212	cluster_1
Row277	111,531,...	391,250,...	cluster_1	0.576	0.212	0.212	cluster_1
Row279	14,446,930	43,557,355	cluster_1	0.576	0.212	0.212	cluster_1
Row280	29,837,722	9,890,282	cluster_1	0.576	0.212	0.212	cluster_1
Row289	16,105,197	13,584,892	cluster_1	0.576	0.212	0.212	cluster_1
Row296	597,624	1,461,186	cluster_1	0.576	0.212	0.212	cluster_1
Row297	311,340	515,291	cluster_1	0.576	0.212	0.212	cluster_1
Row298	67,212	90,433	cluster_1	0.576	0.212	0.212	cluster_1
Row300	745,130	1,277,103	cluster_1	0.576	0.212	0.212	cluster_1
Row312	34,286,540	5,626,485	cluster_1	0.576	0.212	0.212	cluster_1
Row341	15	4	cluster_1	0.576	0.212	0.212	cluster_1
Row343	74,049	8,060	cluster_1	0.576	0.212	0.212	cluster_1
Row351	4,950	16,294	cluster_1	0.576	0.212	0.212	cluster_1
Row358	879,810	1,140,224	cluster_1	0.576	0.212	0.212	cluster_1

4-40 Tabela wyjściowa predykcji

5. Podsumowanie

W trakcie realizacji projektu zapoznaliśmy się z narzędziami wykorzystywanymi w pracy z danymi. Mieliśmy okazję sprawdzić jak sprawdza się pisanie skryptów a jak łączenie węzłów w środowisku graficznym. Nie uniknęliśmy także problemów jak długie wczytywanie dużej ilości rekordów do MySQL czy odmawiająca posłuszeństwa maszyna wirtualna. Zrealizowaliśmy nasze założenia jakimi były zapoznanie się z OLAP oraz użycie algorytmów data mining.

6. Spis ilustracji

7. Źródła

[1] https://docs.knime.com/?category=analytics_platform&version=3.7 dostęp 30.05.2022
[2] https://docs.microsoft.com/pl-PL/sql/integration-services/sql-server-integration-services?view=sql-server-ver16 dostęp 30.05.2022
[3] https://docs.sqlalchemy.org/en/14/ dostęp 30.05.2022
[4] https://druid.apache.org/docs/latest/design/index.html dostęp 30.05.2022
[5] https://dev.mysql.com/doc/ dostęp 30.05.2022