



**WYDZIAŁ
MATEMATYKI
I FIZYKI STOSOWANEJ**
POLITECHNIKI RZESZOWSKIEJ

Wielowymiarowa analiza danych

Projekt

**Gabriel Lichacz
Patryk Motyka**

Rzeszów, 2022

Spis treści

1. Wstęp.....	3
2. ETL – SSIS	3
3. ETL – Python	3
3.1. Skrypt.....	4
4. Knime	7
4.1. ETL	7
4.2. Dashboardy	17
4.3. OLAP	19
4.4. Data mining.....	23
5. Podsumowanie.....	29
6. Spis ilustracji	30
7. Źródła.....	31

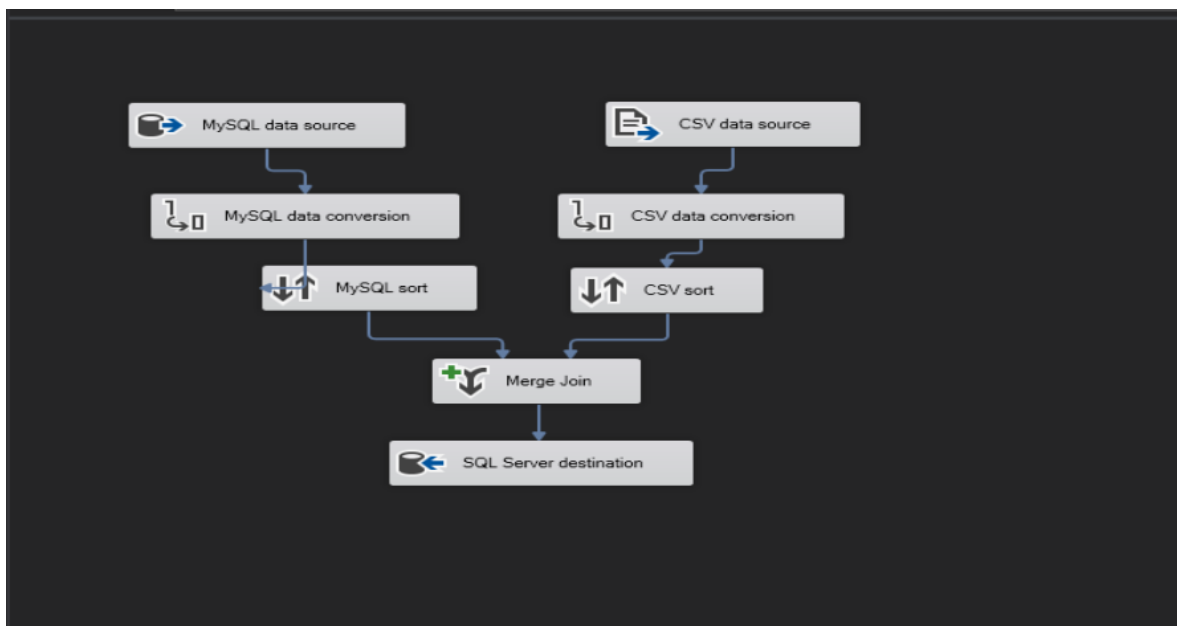
1. Wstęp

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 bazie 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 problemy 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.**

Datasource name	Availability	Availability detail	Total data size	Segment size (rows) minimum / average / maximum	Segment granularity	Total rows	Avg. row size (bytes)	Replicated size	Compaction	% Compacted bytes / segments / intervals	Left to be compacted	Retention	Actions
data	Fully available (1 segment)	No segments to ...	313.52 MB	4.864 M 4.864 M 4.864 M	All	4,863,756	64	313.52 MB	Not enabled	- - -	-	Cluster default: L	🔍 🔗
data_bigger	Fully available (1 segment)	No segments to ...	40.37 MB	0.290 M 0.290 M 0.290 M	All	289,713	139	40.37 MB	Not enabled	- - -	-	Cluster default: L	🔍 🔗
data_smaller	Fully available (1 segment)	No segments to ...	5.29 MB	0.290 M 0.290 M 0.290 M	All	289,652	18	5.29 MB	Not enabled	- - -	-	Cluster default: L	🔍 🔗
telecom_users	Fully available (1 segment)	No segments to ...	568.63 KB	0.006 M 0.006 M 0.006 M	All	5,986	94	568.63 KB	Not enabled	- - -	-	Cluster default: L	🔍 🔗
wikipedia	Fully available (1 segment)	No segments to ...	6.53 MB	0.024 M 0.024 M 0.024 M	Day	24,433	267	6.53 MB	Not enabled	- - -	-	Cluster default: L	🔍 🔗

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 *
```

Połączenie z bazą danych w Apache Druid

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

Kwerendy wczytujące dane

```
# Wczytanie całej 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 ramkę danych pandas
print (df_data_smaller)
```

```
# Wczytanie całej 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')

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

# Usuwanie niepotrzebnej już 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 formatów 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','')
df_data['REASON.FOR.THE.STATUS'] =
```

Wczytanie danych z pliku CSV

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

Czyszczenie danych

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

# Usuwanie zbyt duzej ilosci kolumn (dla szybkości działania 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`);')
```

#	id	country_or_area	year	comm_code	commodity	flow	trade_usd	weight_kg	quantity	category	IMO	SHIPNAME	CLASS	STATUS	REASON_FOR_THE_STATUS
1	1	Cameroon	2012	100630	Rice, semi-milled or wholly milled	Export	383362	720000	720000	10_cereals	9872432	PAVEL LEONOV(10/03/22)	RMRS	Delivered	...
2	2	Cameroon	2012	100640	Rice, broken	Import	9239070	15435351	15432551	10_cereals	9833450	OCEAN PRIME(10/03/22)	ABS	Delivered	...
3	3	Cameroon	2012	100640	Rice, broken	Export	31341	100000	100000	10_cereals	9795969	ORANGE SPIRIT(06/03/22)	BV	Delivered	...
4	4	Cameroon	2012	100700	Grain sorghum	Import	1400027	4081573	4081573	10_cereals	9821574	HMS 6(10/03/22)	ABS	Delivered	...
5	5	Cameroon	2012	100820	Millet	Import	44	11	11	10_cereals	9790854	ULTRA DIVERSITY(11/03/22)	NKK	Delivered	...
6	6	Cameroon	2012	100890	Cereals unmilled nes	Export	858	213	213	10_cereals	9772959	FAST TIGER(10/03/22)	ABS	Delivered	...
7	7	Cameroon	2012	100890	Cereals unmilled nes	Export	24	100	100	10_cereals	9795366	JAVANEAGARA 202(07/07/21)	LRS
8	8	Canada	2012	100110	Durum wheat	Import	14624772	39611528	39611528	10_cereals	9785744	COSCO SHIPPING PEONY(11/03/22)	CCS	Delivered	...
9	9	Canada	2012	100110	Durum wheat	Export	1475691007	3893671470	3893671470	10_cereals	9843194
10	10	Canada	2012	100100	Wheat except durum wheat, and m...	Import	18695625	56962315	56962315	10_cereals	9737656	DE BO 2(15/12/20)	CCS	Delivered	...
11	11	Canada	2012	100190	Wheat except durum wheat, and m...	Export	4675014563	13973149	13973149	10_cereals	9777711	HARVESTER(11/03/22)	LRS	Withdrawn	by society for other re...
12	12	Canada	2012	100190	Wheat except durum wheat, and m...	Re-I...	181161	424000	424000	10_cereals	9749233	AFRICAN GANNET(11/03/22)	NKK	Delivered	...
13	13	Canada	2012	100200	Rye	Import	88421	342345	342345	10_cereals	9813682	CMM CORCORSES(07/03/22)	NV	Delivered	...
14	14	Canada	2012	100200	Rye	Export	53799291	167081966	167081966	10_cereals	9845049	FSL KELANG(06/03/22)	BV	Delivered	...
15	15	Canada	2012	100300	Barley	Import	1044131	4312048	4312048	10_cereals	9853981	CITY - 40(07/03/22)	RMRS	Delivered	...
16	16	Canada	2012	100300	Barley	Export	484739737	1545962956	1545962956	10_cereals	9777723	GOLDEN GRAINS(11/03/22)	LRS	Delivered	...
17	17	Canada	2012	100300	Barley	Re-I...	3121	9000	9000	10_cereals	9754953	MSC DITTE(07/03/22)	NV	Delivered	by society for other re...
18	18	Canada	2012	100400	Oats	Import	2697864	14737389	14737389	10_cereals	9770127	DJANA(02/08/20)	BV	Delivered	...
19	19	Canada	2012	100400	Oats	Export	426994782	1643538365	1643538365	10_cereals	9903865	POLA MARINA(10/03/22)	RMRS	Delivered	...
20	20	Canada	2012	100400	Oats	Re-I...	87	256	256	10_cereals	9851270	NORA B(06/03/22)	BV	Delivered	...
21	21	Canada	2012	100510	Maize (corn) seed	Re-I...	4243895	686000	686000	10_cereals	9755866	AVRA GR(07/03/22)	RMRS	Delivered	...
22	22	Canada	2012	100510	Maize (corn) seed	Import	147875270	31229943	31229943	10_cereals	9803349	GIELO DI IVOL(11/12/21)	NKK	Delivered	...
23	23	Canada	2012	100510	Maize (corn) seed	Export	96746619	22668473	22668473	10_cereals	9839911	AM TARANG(07/03/22)	NV	Delivered	...

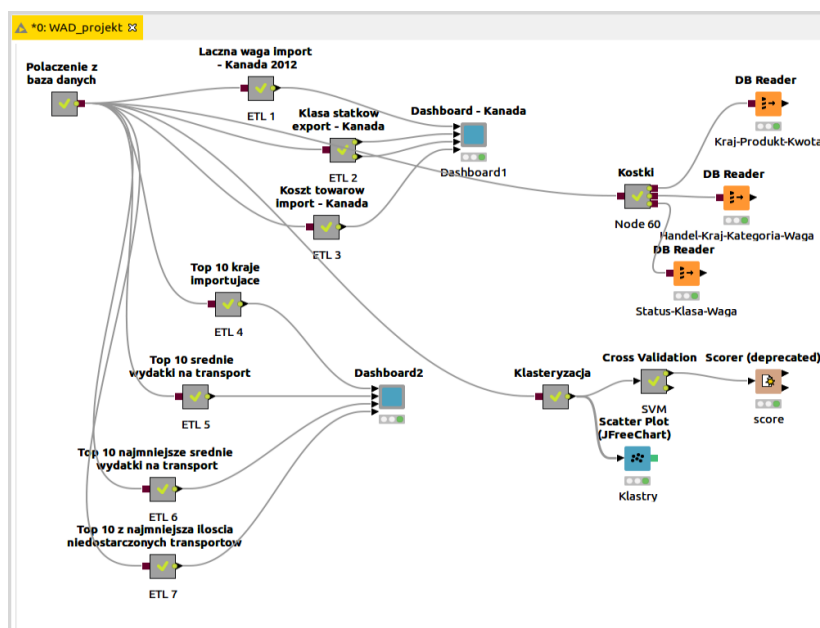
3-2 Dane w MySQL po wykonanym procesie ETL

4. Knime

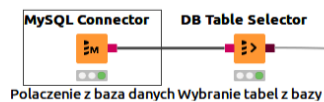
Knime to platforma przeznaczona do integracji oraz analizy, a także raportowania danych. Operacje przeprowadza się za pomocą zestawiania węzłów (nodes), w których można także korzystać z języków jak np. SQL. Knime jest dostępny na zasadzie otwartego oprogramowania.

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



Dialog - 0:30:1 - MySQL Connector (Połączenie z baza danych)

File

Input Type Mapping Output Type Mapping Flow Variables

Connection Settings JDBC Parameters Advanced

Configuration

Database Dialect: MySQL

Driver Name: MySQL 8 [ID: MySQL 8]

Location

Hostname: localhost Port: 3,306

Database name: shipment

Authentication

☐ Credentials

☒ Username & password

Username: mysql_admin

Password:

OK Apply Cancel ?

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

Dialog - 0:1 - MySQL Connector

File

Input Type Mapping Output Type Mapping Flow Variables

Connection Settings JDBC Parameters Advanced

Connection

Name	Value
Automatically reconnect to database	<input type="checkbox"/>
Reconnect to database timeout	0
Restore database connection	<input type="checkbox"/>
Validation query	SELECT 0

Dialect syntax

Name	Value
Delimit only identifier with spaces	<input type="checkbox"/>
Identifier delimiter (closing)	`
Identifier delimiter (opening)	`

JDBC logger

Name	Value
Enable	<input type="checkbox"/>

JDBC statement cancellation

Name	Value
Enable	<input checked="" type="checkbox"/>
Node cancellation polling interval	1000

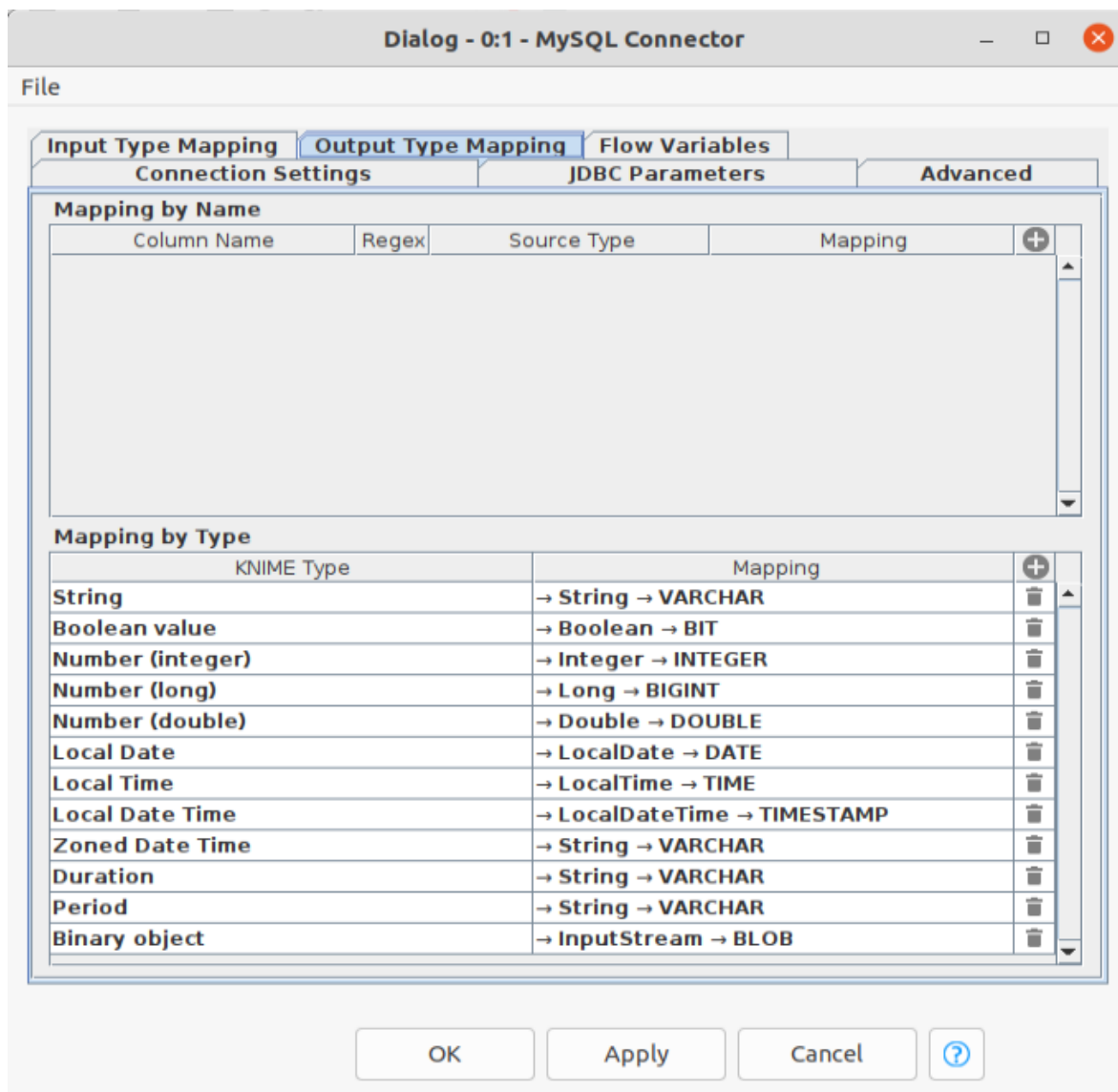
Metadata

Name	Value
Retrieve in configure	<input checked="" type="checkbox"/>
Retrieve in configure timeout	3

Name	Value
Fail if WHERE clause contains any missing value	<input checked="" type="checkbox"/>
Fetch size	10000

OK Apply Cancel ?

4-3 Połączenie Knime z bazą w mysql - Advanced



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

Dialog - 0:1 - MySQL Connector

File

Input Type Mapping Output Type Mapping **Flow Variables**

Connection Settings JDBC Parameters Advanced

? authentication

- s credentials
- s username
- s password
- s selectedType

? session_info

? mysql-connection

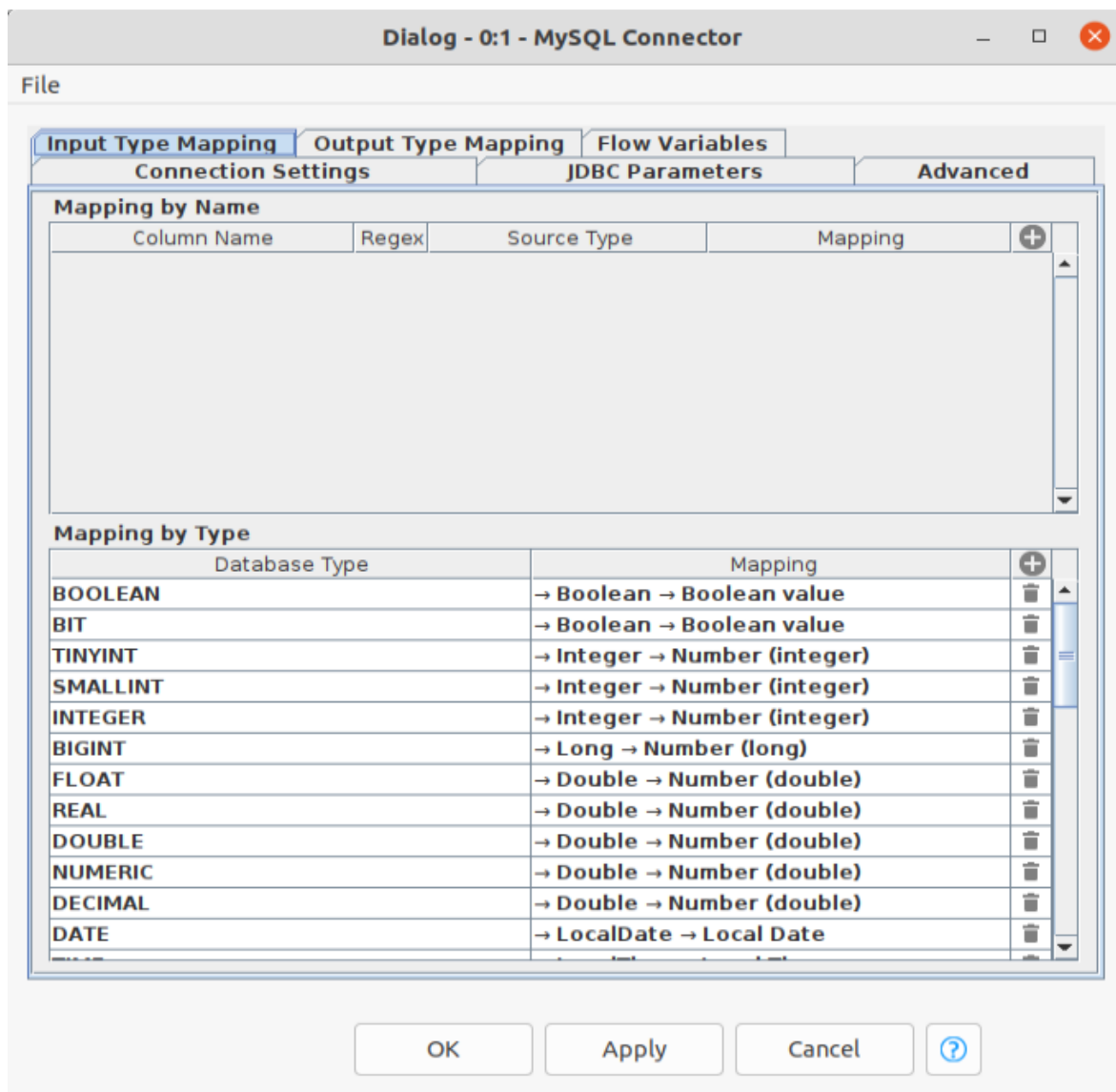
- s host
- i port
- s database_name

? external_to_knime_mapping

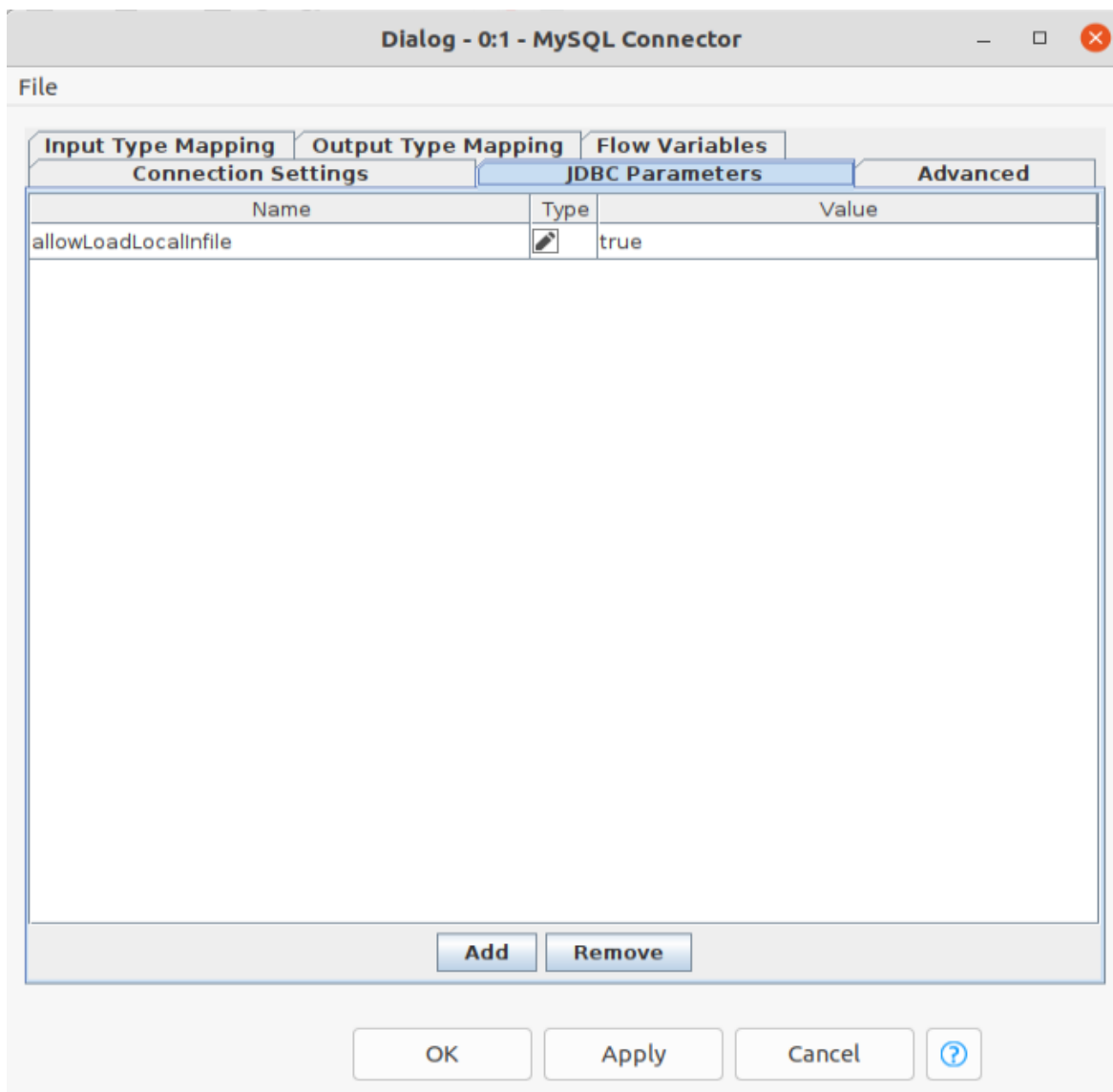
? knime_to_external_mapping

OK Apply Cancel ?

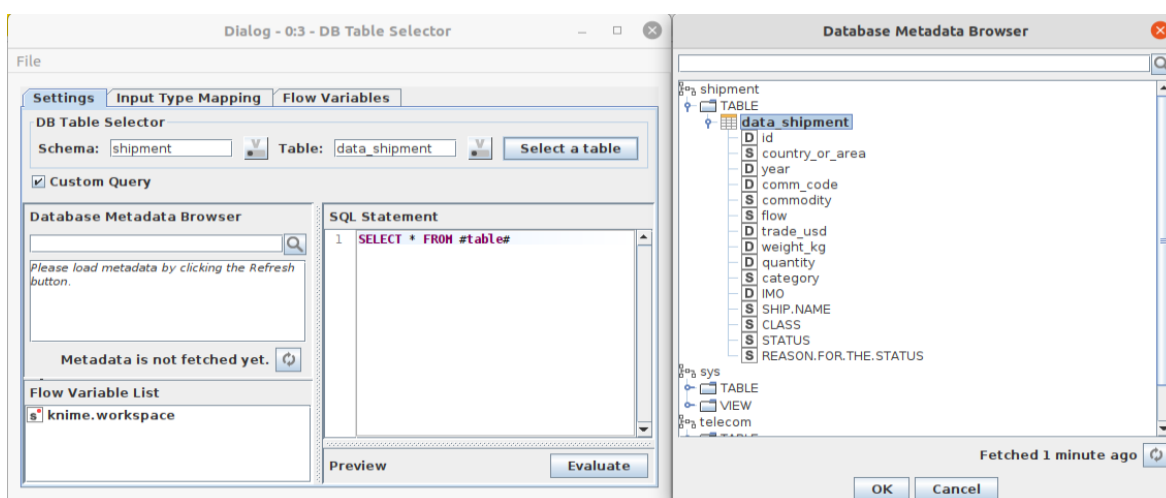
4-5 Połączenie Knime z bazą w mysql - Flow Variables



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

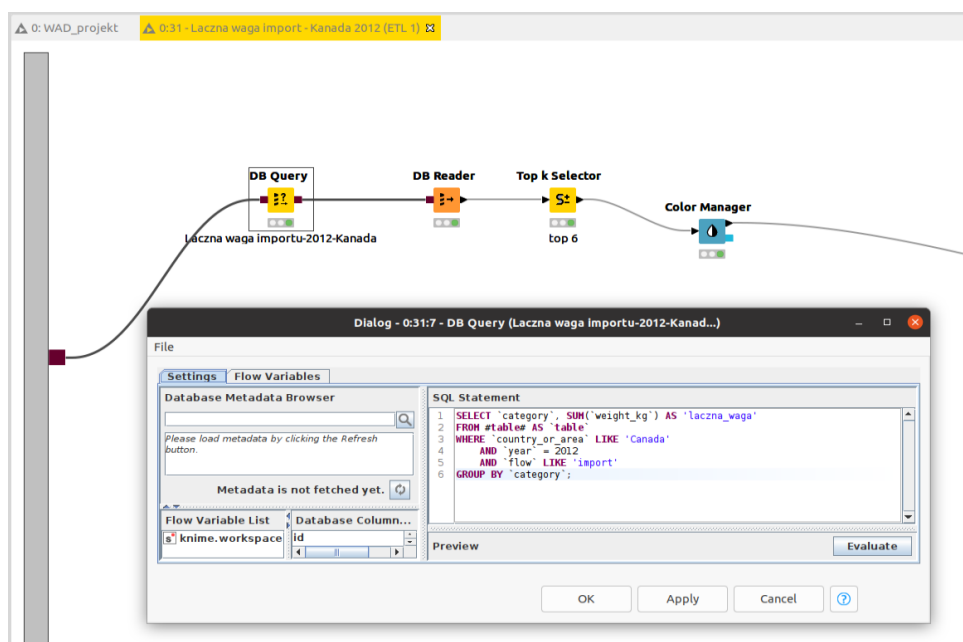


4-7 Połączenie Knime z bazą w mysql - JDBC Parameters

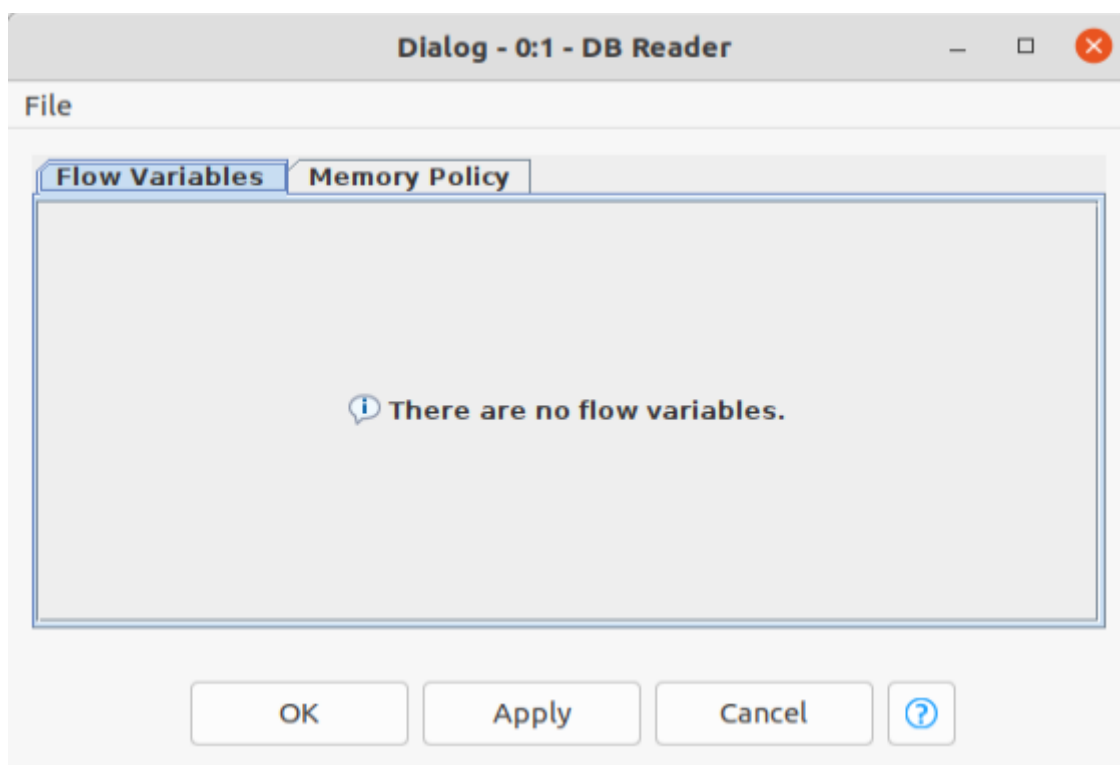


4-8 Table Selector – Settings

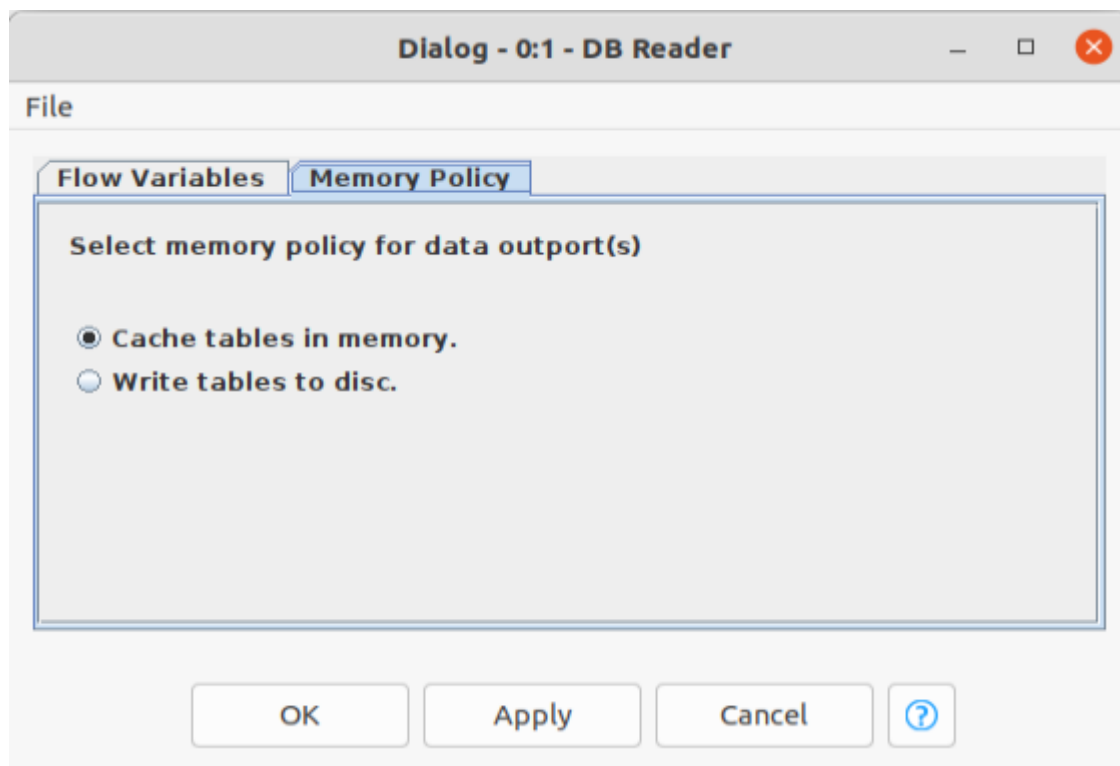
Poniżej znajduje się kilka z ETL, które wykonaliśmy przed tworzeniem wykresów. Pierwsza część dotyczy Kanady, druga natomiast porównania różnych krajów.



4-9 ETL z kanadyjskim importem w 2012 roku

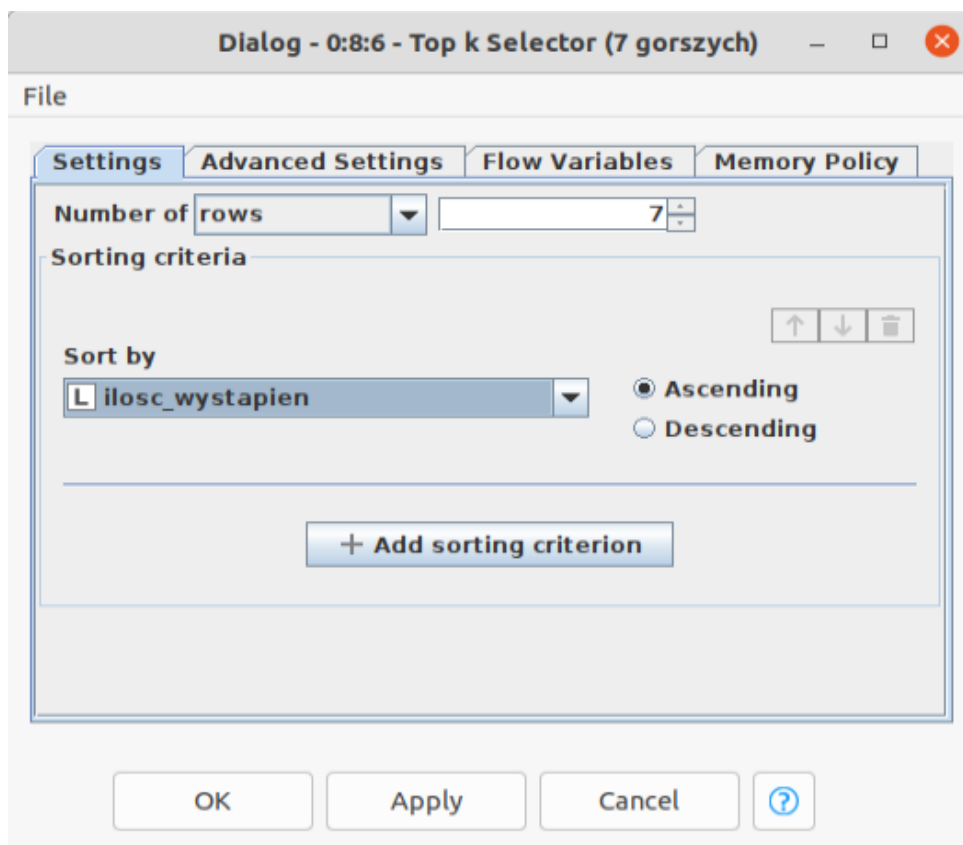


4-10 Okno edycji w węźle DB Reader (Flow Variables)

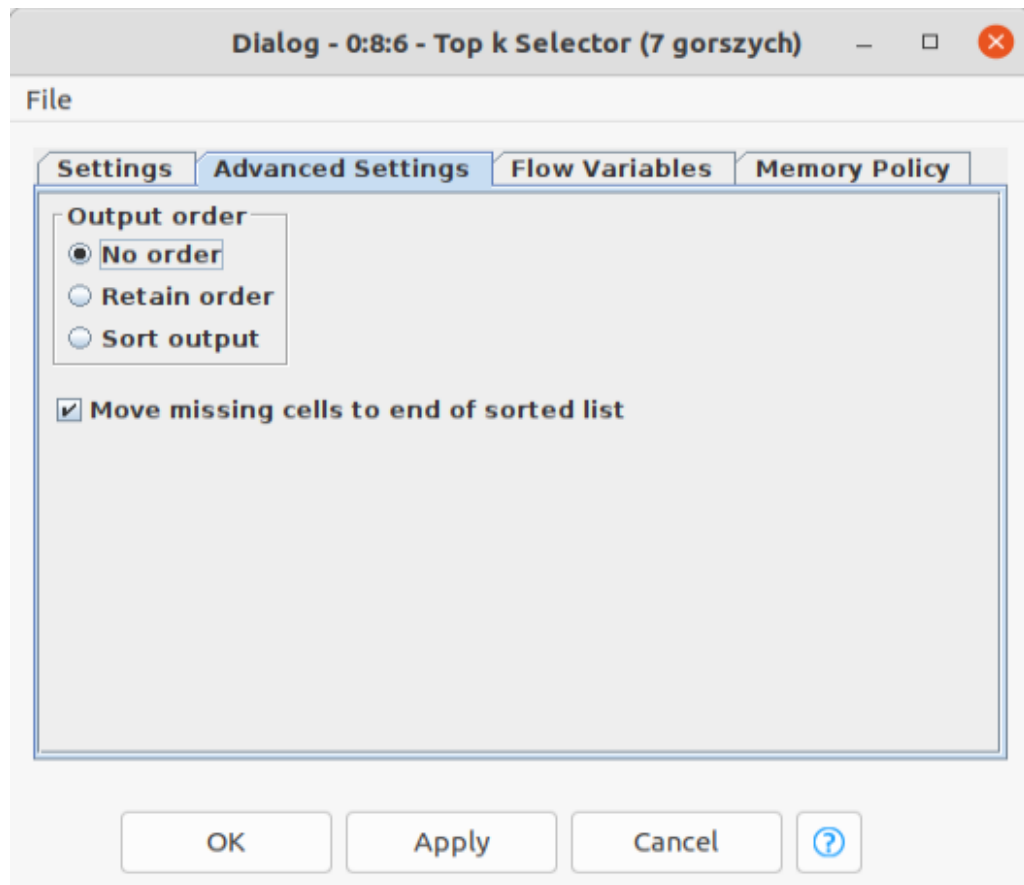


4-11 Okno edycji w węźle DB Reader (Memory Policy)

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

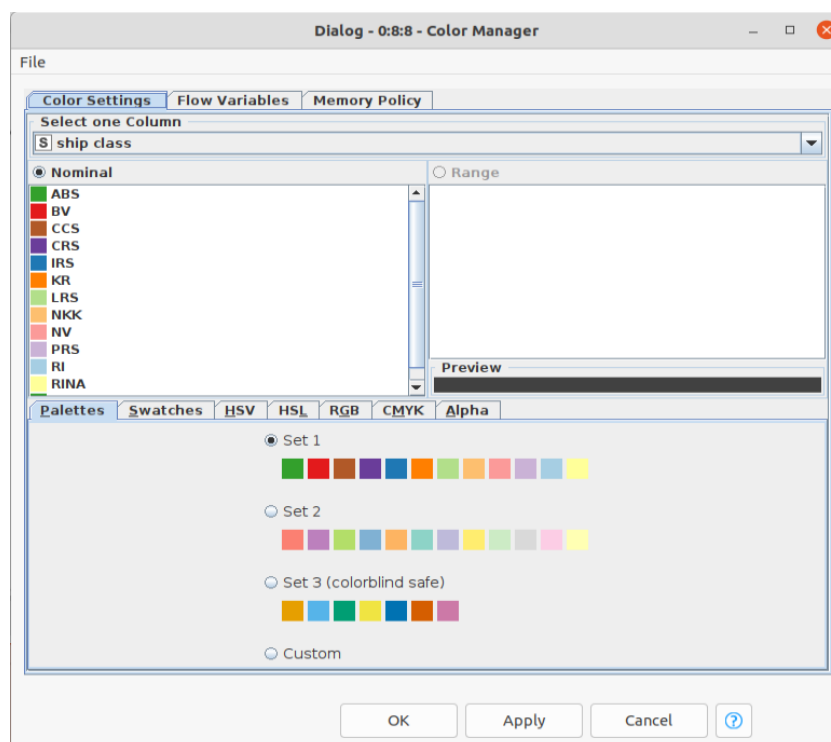


4-12 Top K Selector – Settings

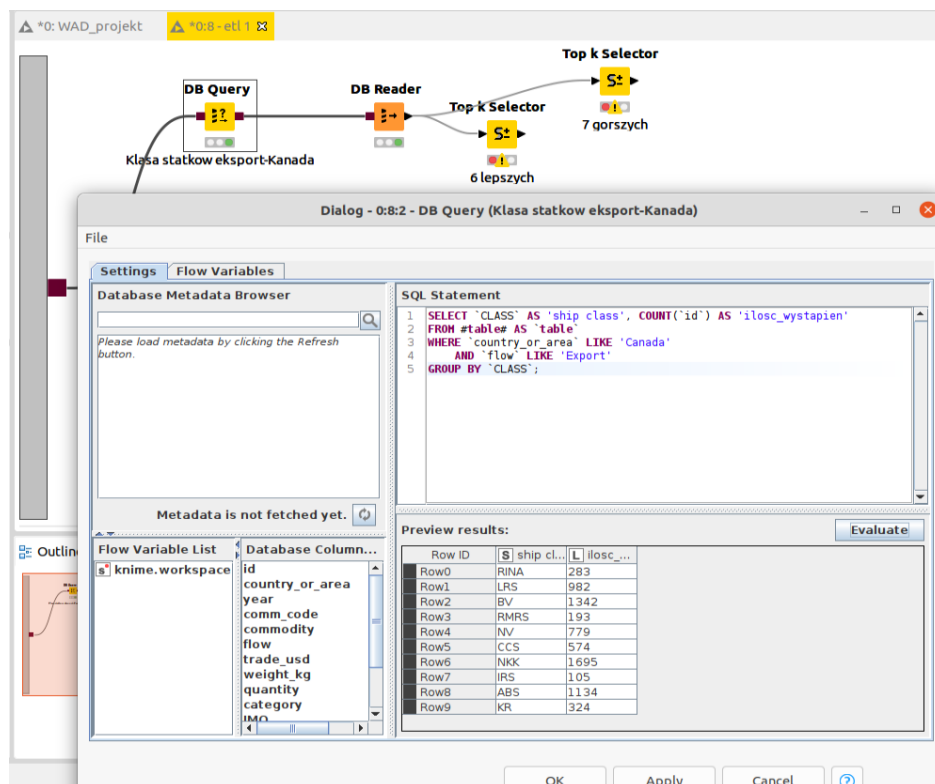


4-13 Top K Selector - Advanced Settings

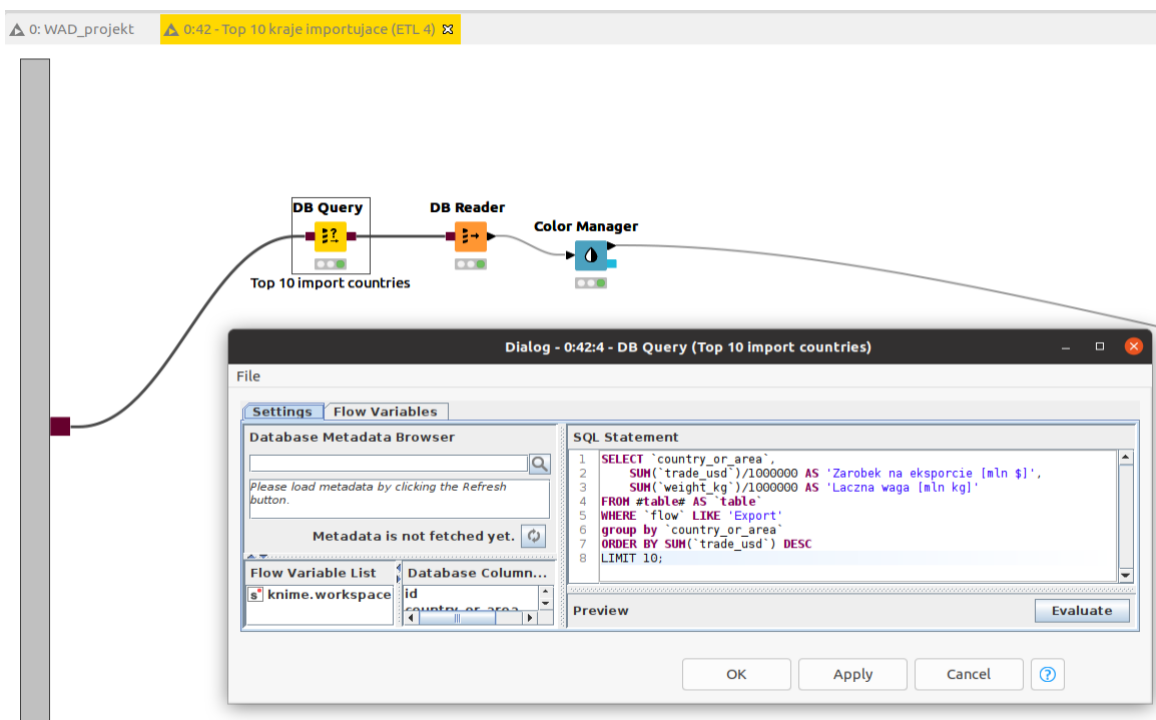
Top K Selector służy do ograniczenia liczby zwracanych wyników.



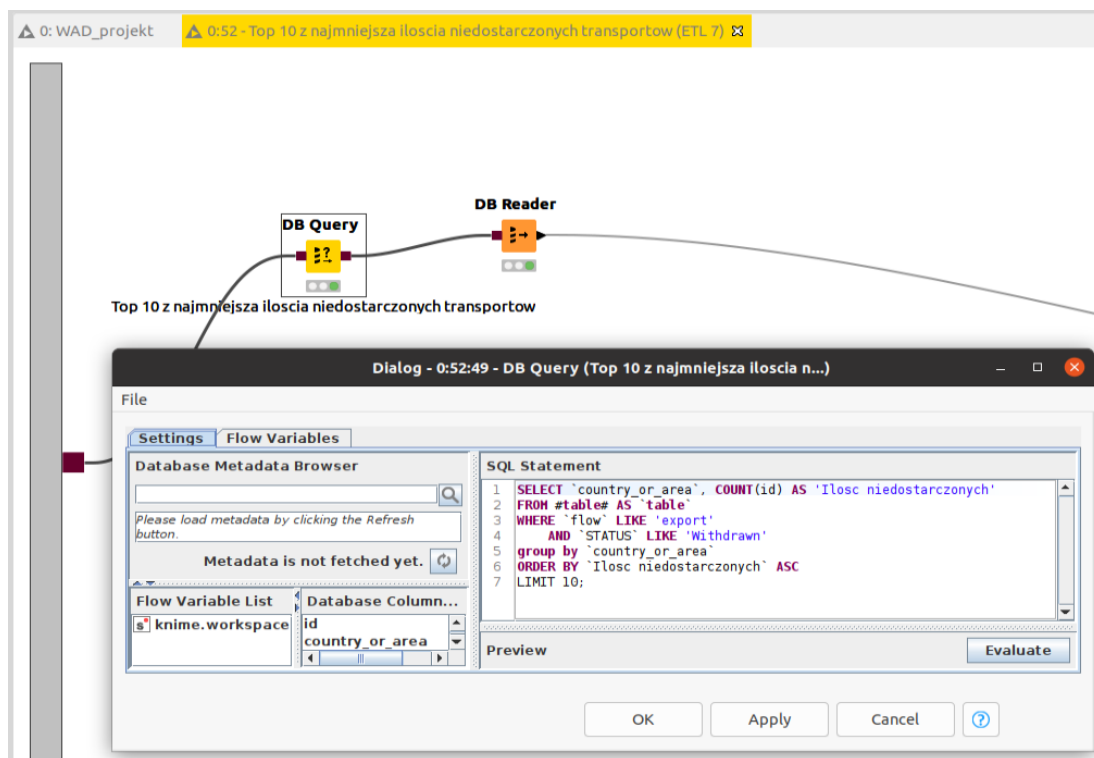
4-14 Color Manager



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



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

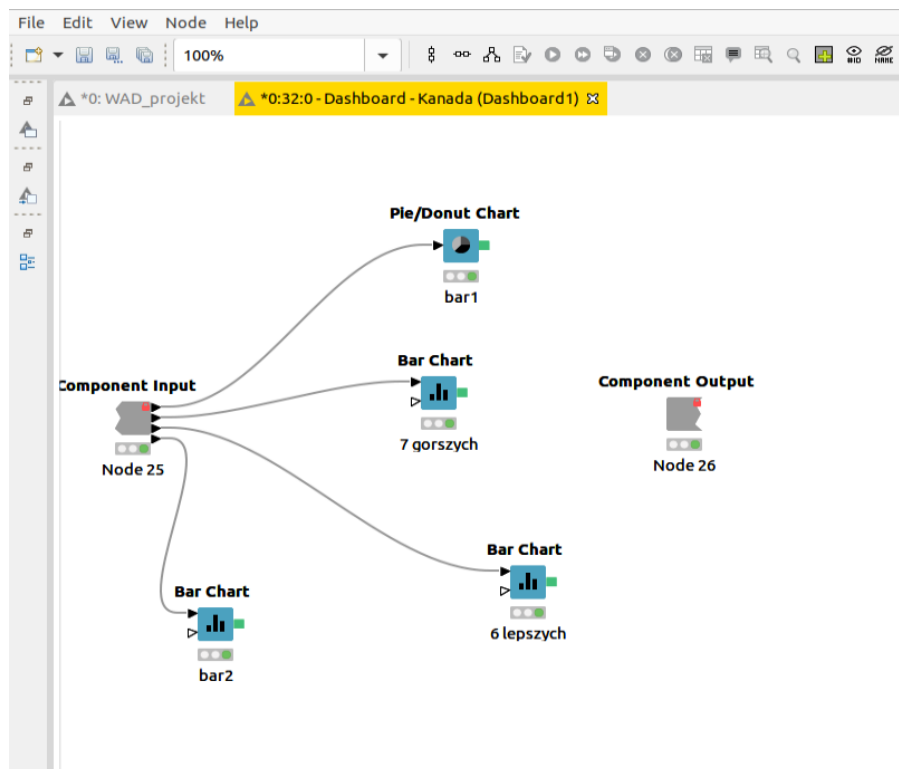


4-17 ETL z krajami, które mają najmniej niedostarczonych transportów

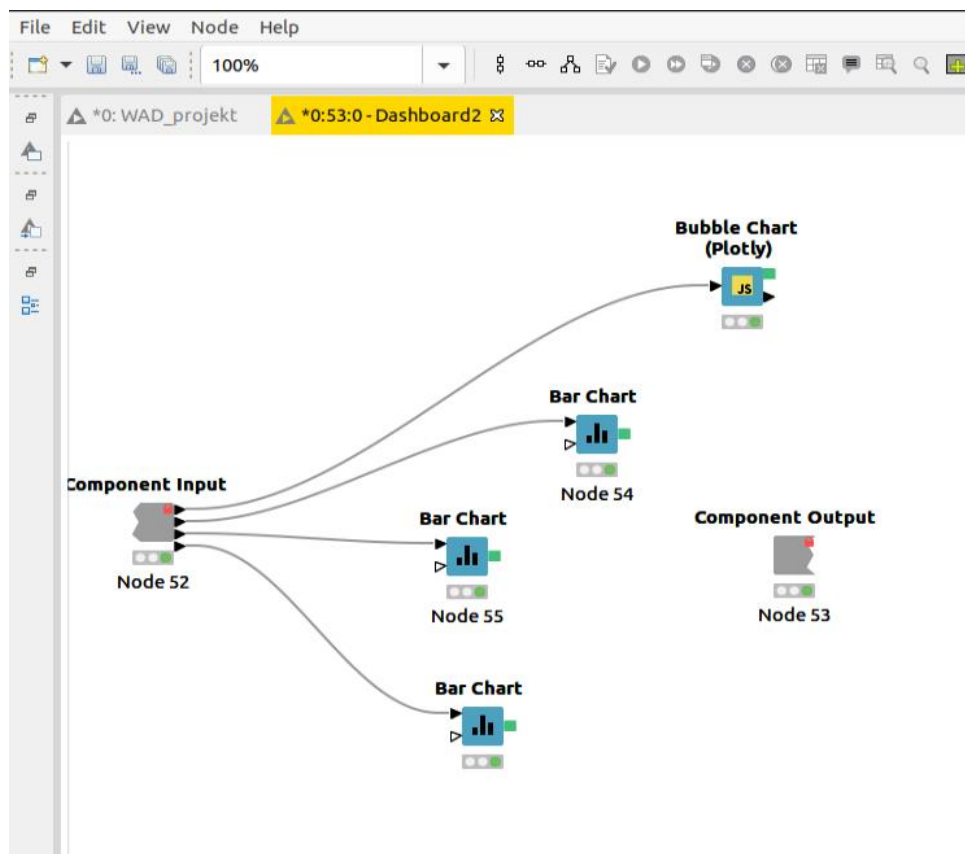
Jak można zauważyć, często korzystaliśmy po prostu z kwerend pisanych w SQL, ponieważ nasze dane były już na tyle przygotowane, że było to narzędzie wystarczające.

4.2. Dashboardy

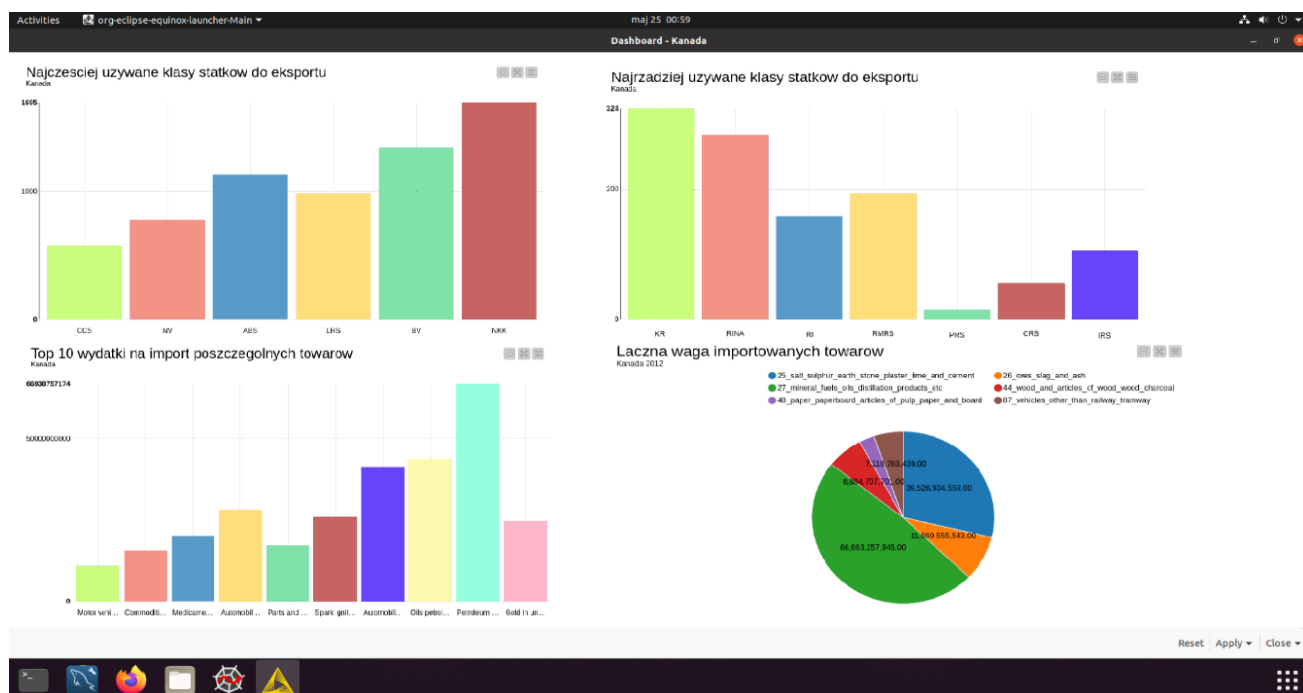
Poniżej znajdują się dwa dashboardy, poprzedzone widokiem ich wykonania w Knime.



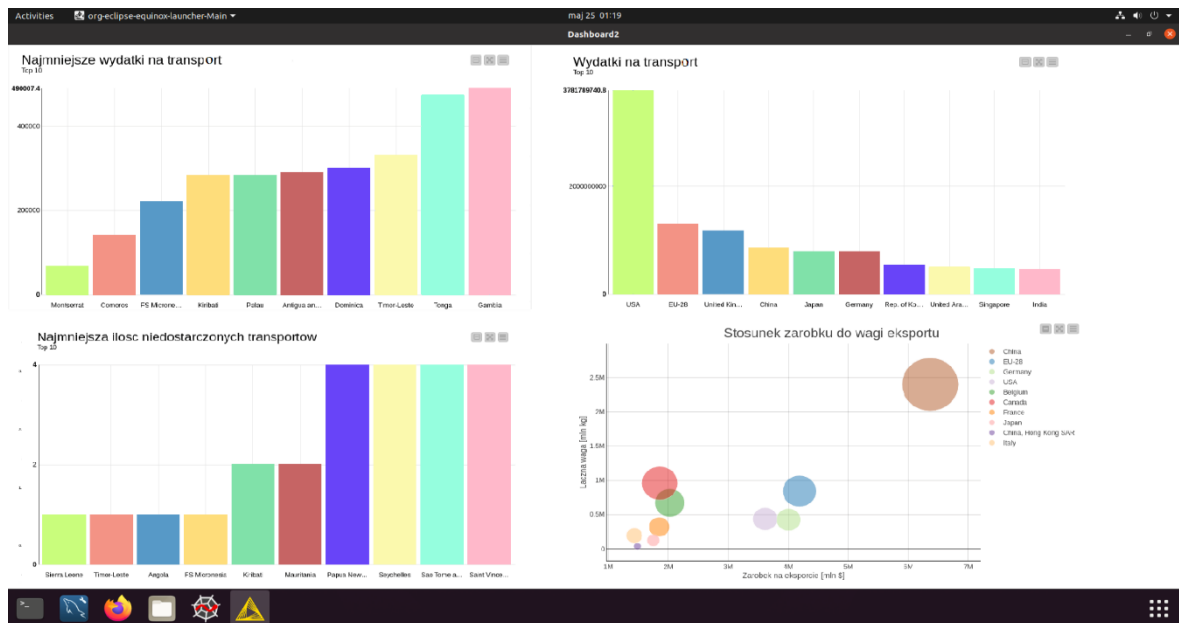
4-18 Budowa dashboardu dla Kanady



4-19 Budowa dashboardu z porównaniem krajów



4-20 Dashboard dla Kanady



4-21 Dashboard z porównaniem krajów

Dashboard dla Kanady odpowiada na pytania:

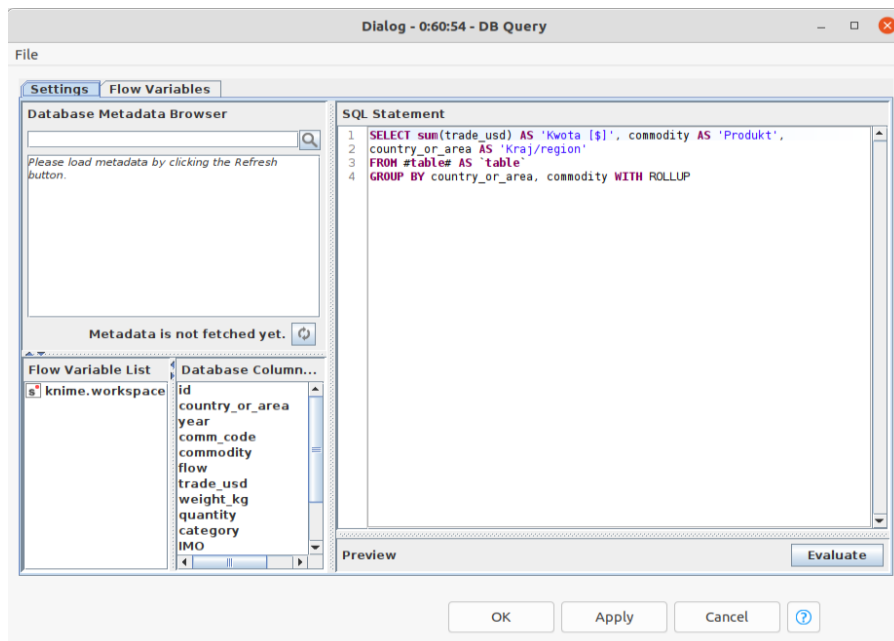
- Jakie klasy statków są najczęściej używane do eksportu w Kanadzie?
- Jakie klasy statków są najrzadziej używane do eksportu w Kanadzie?
- Na import jakich towarów Kanada wydaje najwięcej?
- Towarów których kategorii najwięcej importowała Kanada w 2012 roku?

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

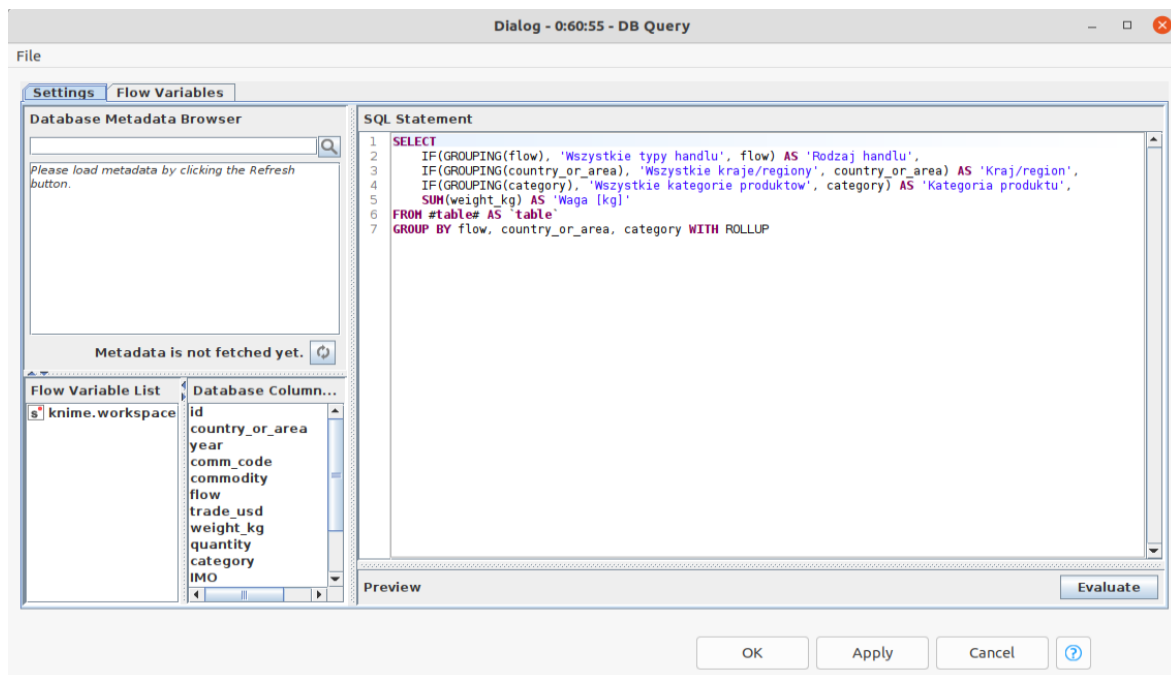
Stworzyliśmy także trzy kostki OLAP., korzystając między innymi z klauzuli WITH ROLLUP. Kostki OLAP są to struktury danych, które pozwalają na szybką analizę danych. Przechowują one dane w sposób bardziej przypominający wielowymiarowe arkusze kalkulacyjne niż tradycyjną, relacyjną bazę danych



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 Hilitte 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 powder or 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 water	Albania
Row127	5,517,339	Acrylic or vinyl polymer paint or varnish, non-aqueou	Albania

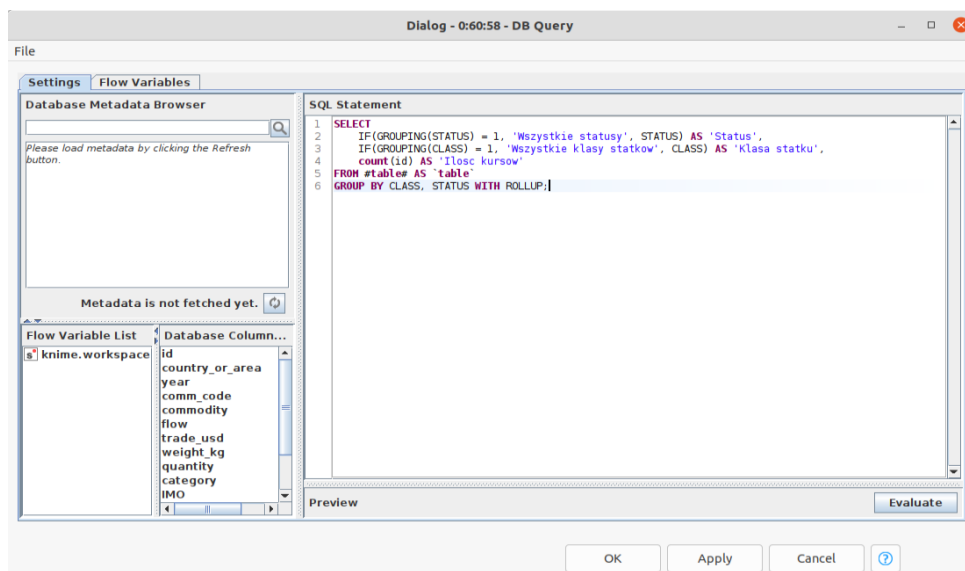
4-23 Widok kostki z podsumowaniem przepływu gotówki w handlu wszystkimi towarami..



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	S Rodzaj handlu	S Kraj/region	S Kategoria produktu	D Waga [kg]
Row26097	Re-Import	United Kingdom	10 cereals	385,309
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,...

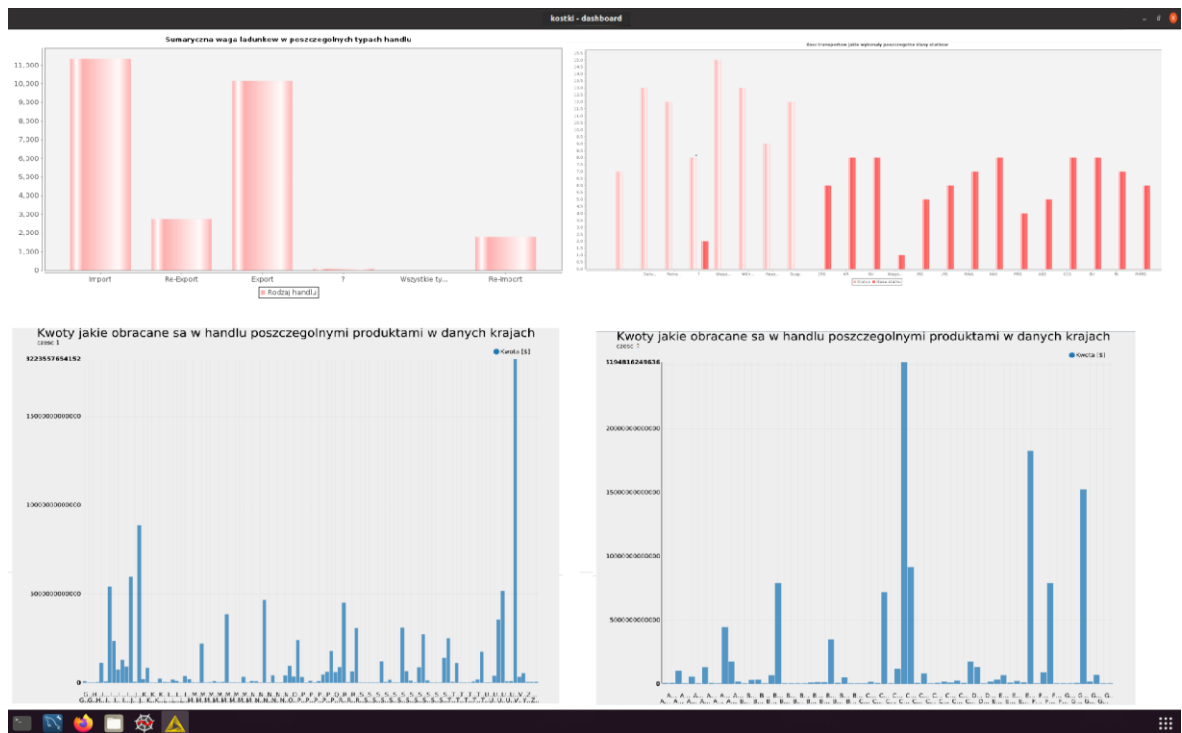
4-25 Kostka na temat handlu w poszczególnych krajach



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

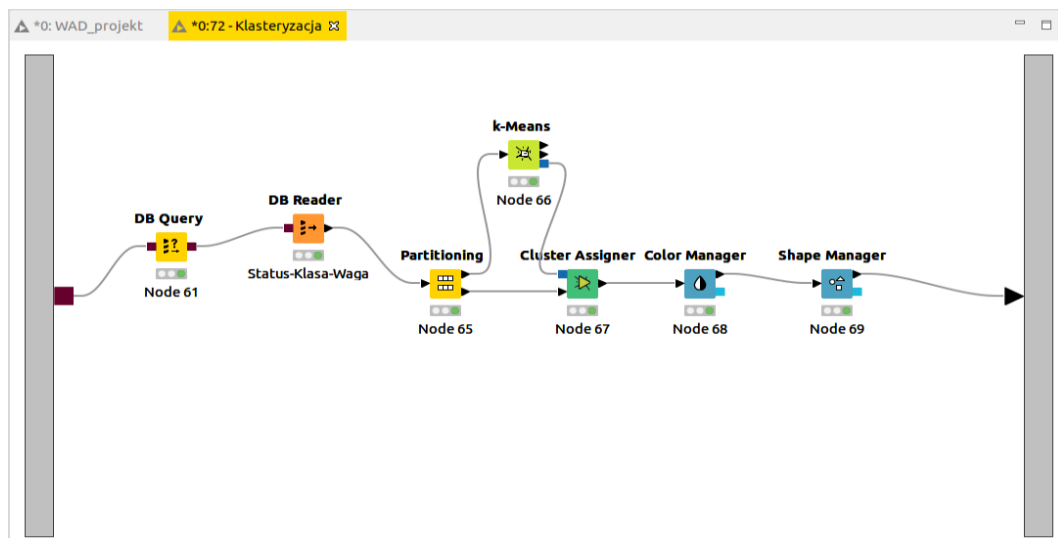
4-27 Kostka z informacjami na temat transportów i wykorzystywanych statków



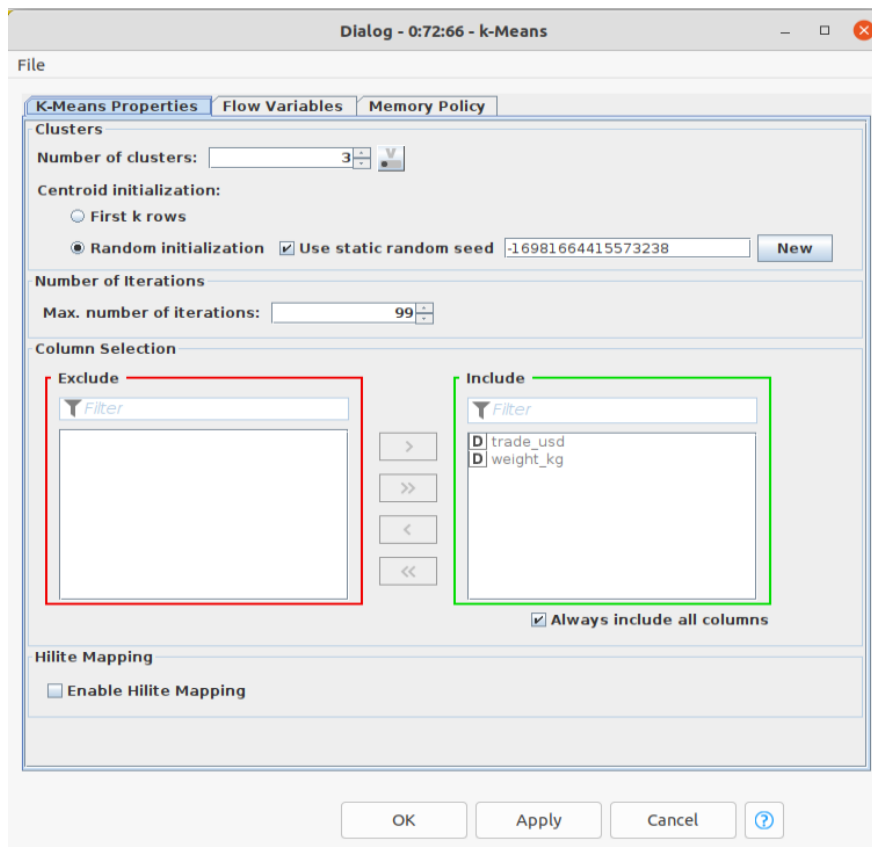
4-28 Dashboard zrobiony na podstawie kostek OLAP

4.4. Data mining

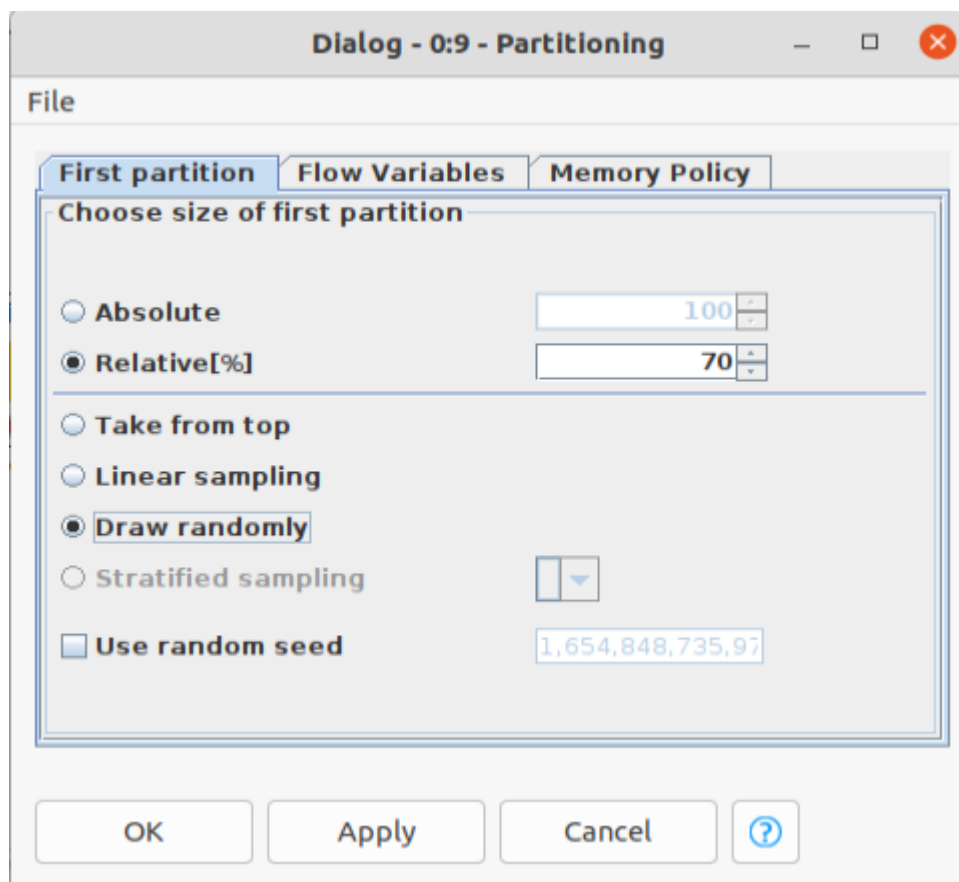
Wykorzystaliśmy także klasteryzację k-Means jako jeden z elementów data mining.



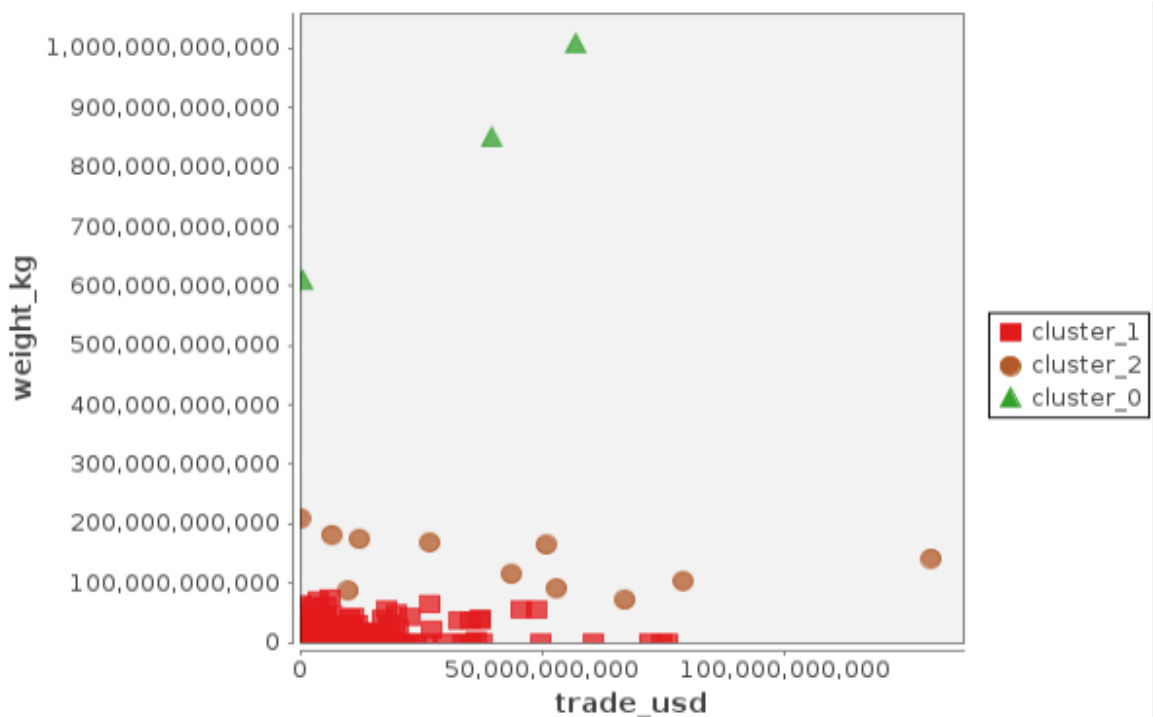
4-29 Budowa klasteryzacji



4-30 Ustawienie parametrów dla algorytmu k-Means

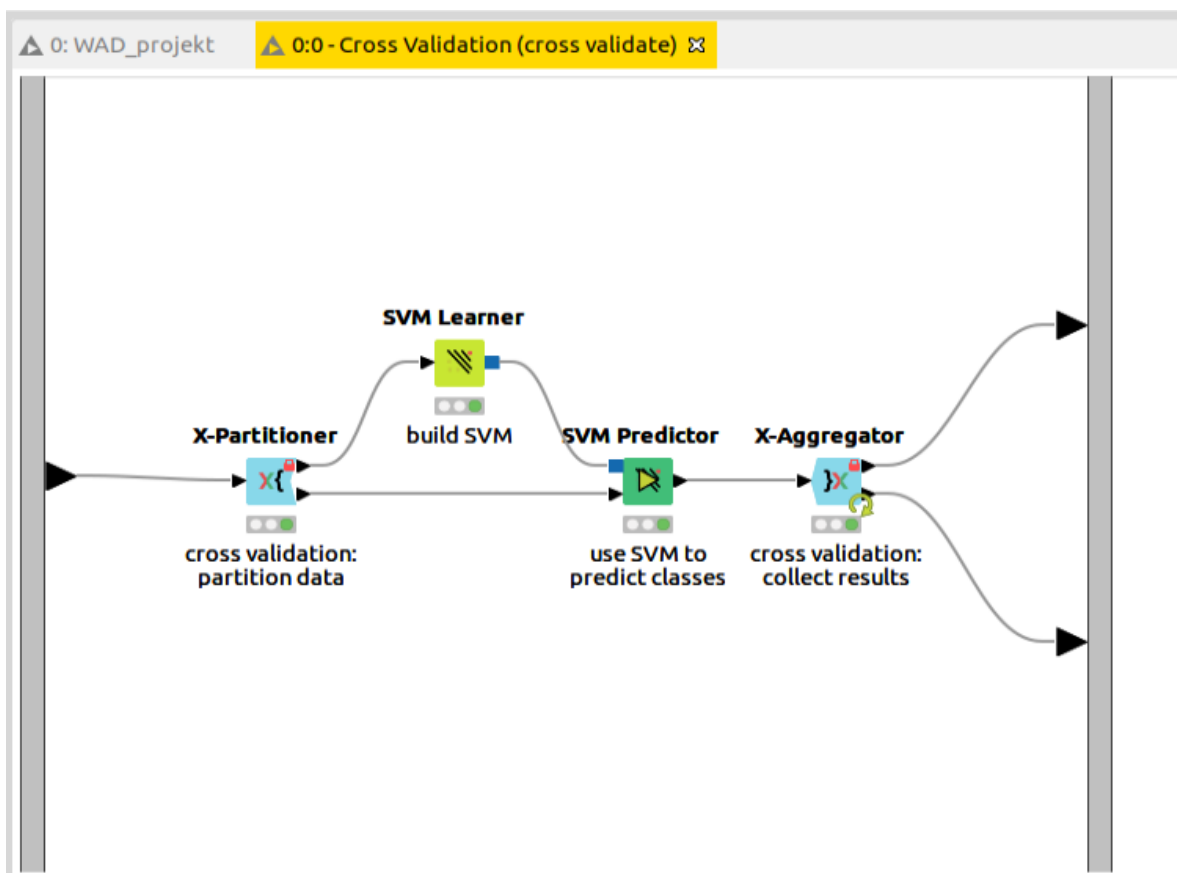


4-31 Wnętrze węzła Partitioning

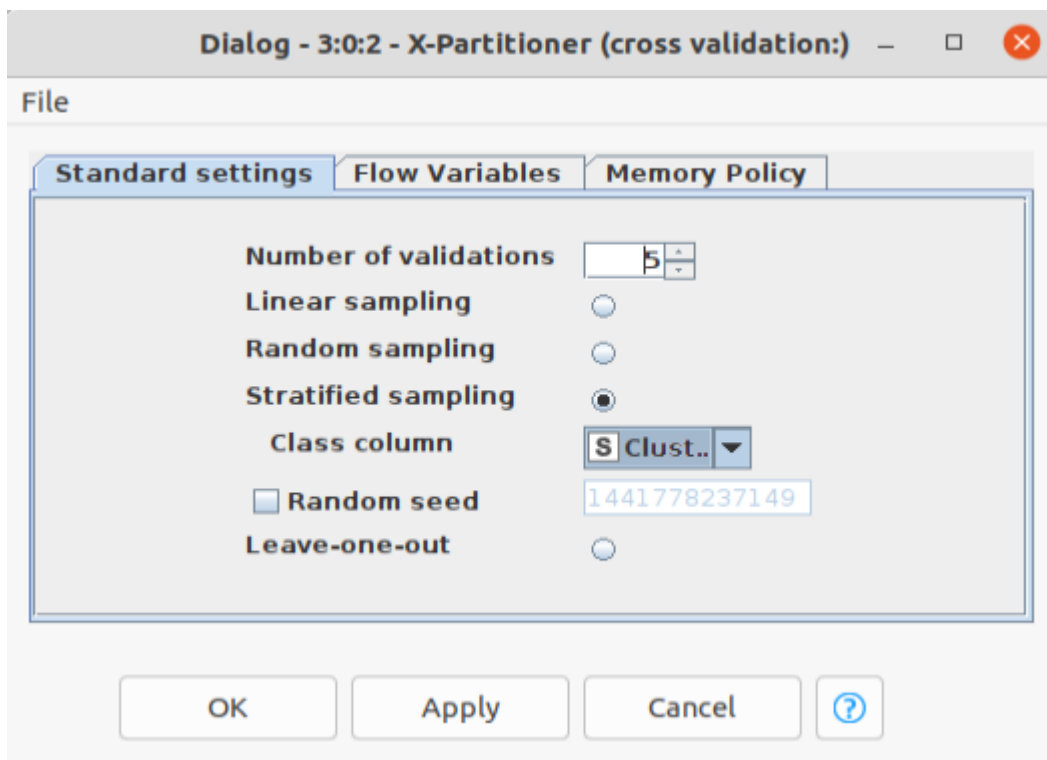


4-32 Wykres przedstawiający wyniki klasteryzacji

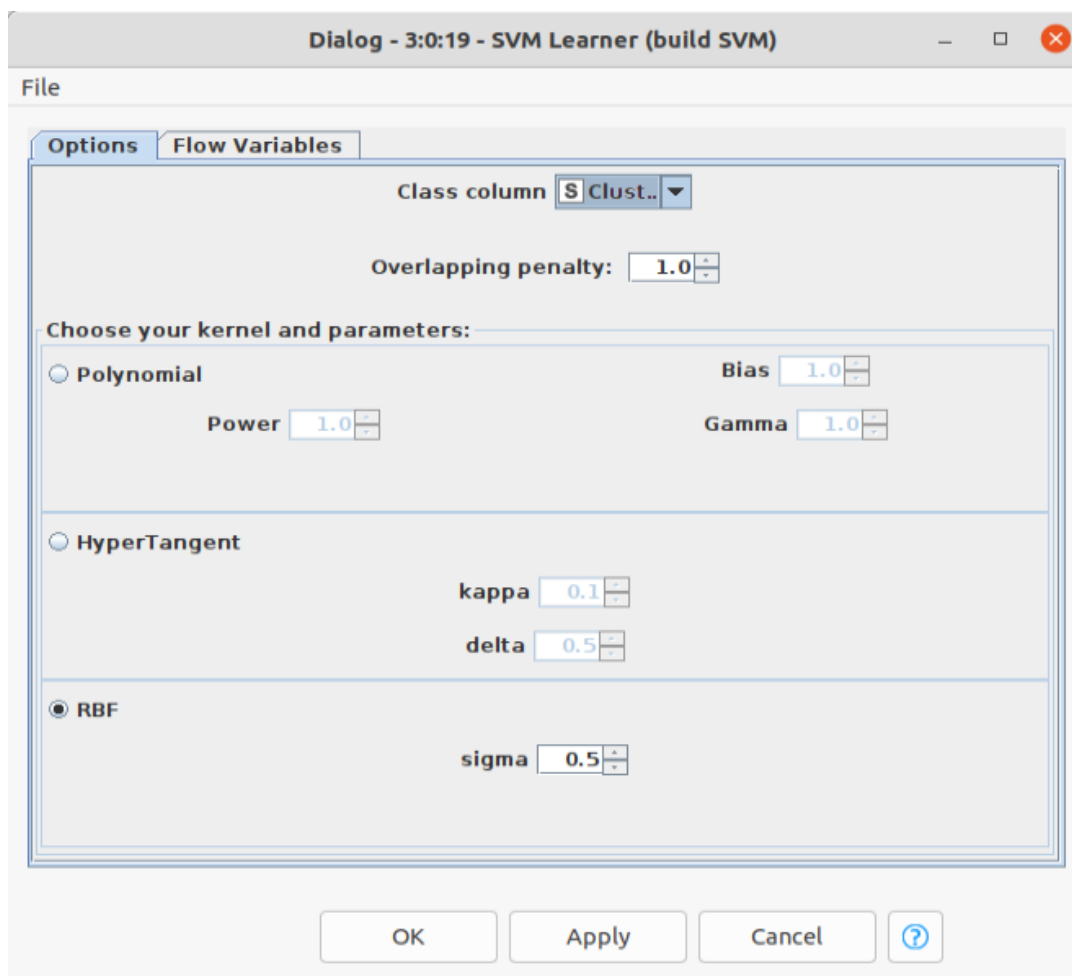
Większość danych została sklasyfikowana do pierwszego klastra, oznaczonego poprzez czerwone kwadraty. Na danych po klasteryzacji wykorzystaliśmy algorytm wektorów wspierających SVM.



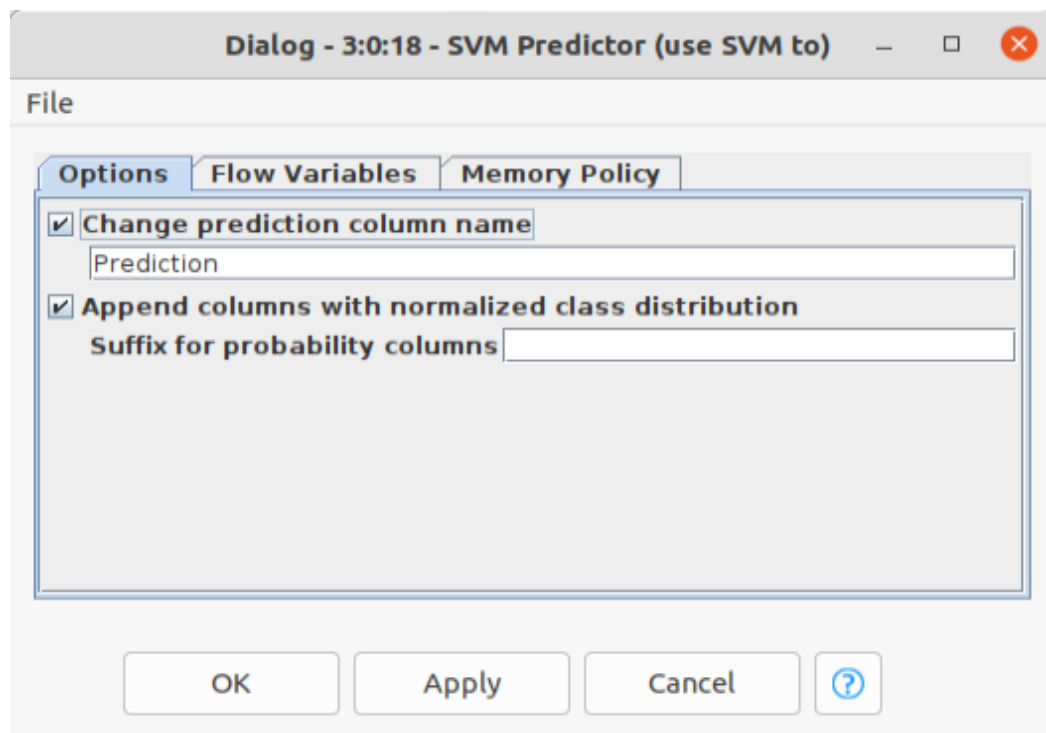
4-33 Budowa modelu SVM



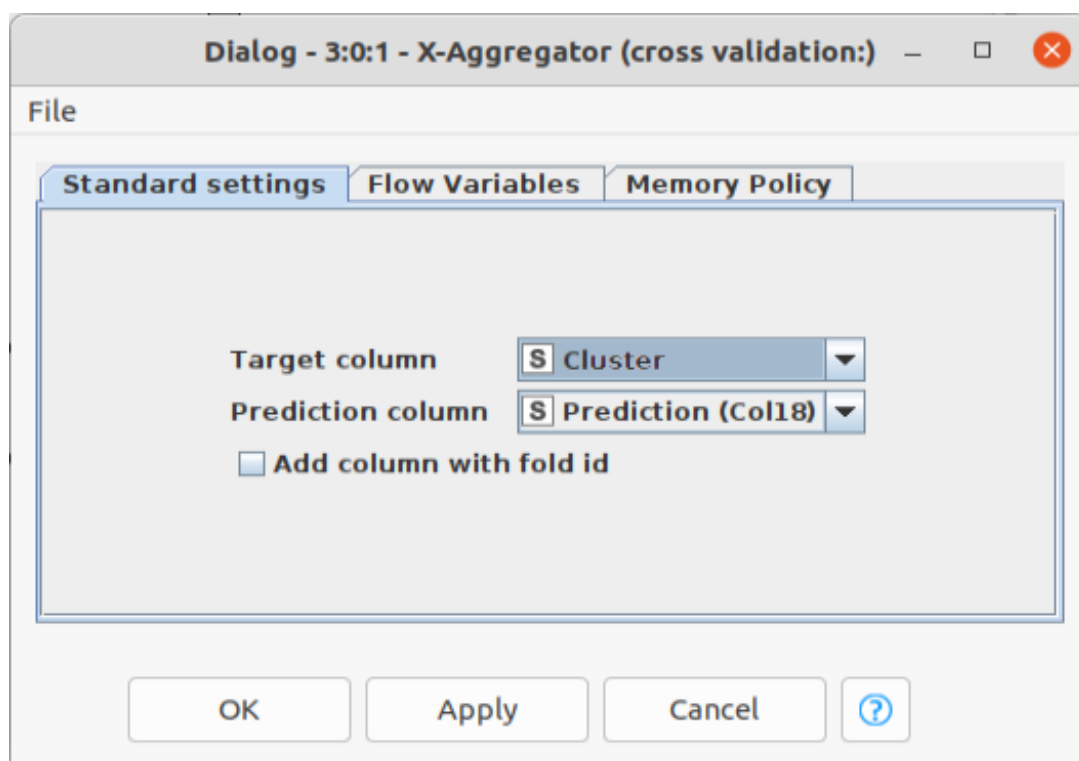
4-34 Wnętrze węzła X-Partitioner służącego do podziału danych walidacji krzyżowej



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	I cluster_1	I cluster_2	I 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	D Error in %	I Size of Test Set	I 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	D trade...	D weight...	S Cluster	D P (Cluster=cluster_1)	D P (Cluster=cluster_2)	D P (Cluster=cluster_0)	S 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. Mieliliśmy 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

2-1 Przepływ danych, który mieliśmy przed problemami z maszyną	3
3-1 Tabele z danymi w Druid	4
3-2 Dane w MySQL po wykonanym procesie ETL	7
4-1 Workspace w Knime.....	7
4-2 Połączenie Knime z bazą w mysql - Connection Settings.....	8
4-3 Połączenie Knime z bazą w mysql - Advanced.....	8
4-4 Połączenie Knime z bazą w mysql - Output Type Mapping	9
4-5 Połączenie Knime z bazą w mysql - Flow Variables	10
4-6 Połączenie Knime z bazą w mysql - Input Type Mapping	11
4-7 Połączenie Knime z bazą w mysql - JDBC Parameters	12
4-8 Table Selector – Settings	12
4-9 ETL z kanadyjskim importem w 2012 roku	13
4-10 Okno edycji w węźle DB Reader (Flow Variables)	13
4-11 Okno edycji w węźle DB Reader (Memory Policy).....	14
4-12 Top K Selector – Settings	14
4-13 Top K Selector - Advanced Settings	15
4-14 Color Manager	15
4-15 ETL z użyciem poszczególnych klas statków przez Kanadę w eksporcie	16
4-16 ETL z zarobkiem i wagą eksportu w poszczególnych krajach.....	16
4-17 ETL z krajami, które mają najmniej niedostarczonych transportów	17
4-18 Budowa dashboardu dla Kanady	17
4-19 Budowa dashboardu z porównaniem krajów.....	18
4-20 Dashboard dla Kanady.....	18
4-21 Dashboard z porównaniem krajów	19
4-22 Kwerenda do stworzenia kostki z kwotą jaka obracana jest w handlu poszczególnymi produktami w danych krajach	20
4-23 Widok kostki z podsumowaniem przepływu gotówki w handlu wszystkimi towarami..	20
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	21
4-25 Kostka na temat handlu w poszczególnych krajach	21
4-26 Kwerenda do stworzenia kostki z informacjami na temat transportów w kontekście klas statków	22
4-27 Kostka z informacjami na temat transportów i wykorzystywanych statków	22
4-28 Dashboard zrobiony na podstawie kostek OLAP.....	23
4-29 Budowa klasteryzacji.....	23
4-30 Ustawienie parametrów dla algorytmu k-Means	24
4-31 Wnętrze węzła Partitioning.....	24
4-32 Wykres przedstawiający wyniki klasteryzacji.....	25
4-33 Budowa modelu SVM	25
4-34 Wnętrze węzła X-Partitioner służącego do podziału danych walidacji krzyżowej.....	26
4-35 Wnętrze węzła SVM Learner	26
4-36 Wnętrze węzła SVM Predictor	27

4-37 Wnętrze węzła X-Aggregator służącego do zebrania wyników w walidacji krzyżowej	27
4-38 Przyporządkowanie	28
4-39 Zliczanie błędów	28
4-40 Tabela wyjściowa predykcji	28

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