



Fundusze  
Europejskie  
Wiedza Edukacja Rozwój



Politechnika Wrocławska

Unia Europejska  
Europejski Fundusz Społeczny



*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Hurtownie danych

## Wprowadzenie

**dr inż. Marcin Maleszka**

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

## **WYMAGANIA WSTĘPNE W ZAKRESIE WIEDZY, UMIEJĘTNOŚCI I INNYCH KOMPETENCJI**

- Posiadanie wiedzy w zakresie organizacji systemów bazodanowych ze szczególnym uwzględnieniem modelu relacyjnego
- Podstawowa znajomość języka zapytań SQL

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

## CELE PRZEDMIOTU

- Opanowanie podstawowej wiedzy i umiejętności posługiwania się operatorami grupującymi SQL oraz funkcjami agregującymi i grupującymi SQL
- Opanowanie podstawowej wiedzy i umiejętności dotyczącej charakterystyk przetwarzania zorientowanego na transakcje (OLTP) oraz przetwarzania zorientowanego na analizę (OLAP)
- Opanowanie podstawowej wiedzy oraz umiejętności posługiwania się hurtownią danych
- Zapoznanie się ze środowiskiem pracy MS PowerPivot, MS SQL Analysis Services, MS SQL Integration Services oraz MS SQL Reporting Services
- Opanowanie podstawowej wiedzy i umiejętności dotyczącej integracji, raportowania oraz wizualizacji danych

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# PRZEDMIOTOWE EFEKTY KSZTAŁCENIA

Z zakresu wiedzy:

- student ma podstawową wiedzę związaną z zastosowaniem i organizacją hurtowni danych
- student ma podstawową wiedzę związaną z procesem ETL, raportowaniem oraz analizą danych



*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# PRZEDMIOTOWE EFEKTY KSZTAŁCENIA

Z zakresu umiejętności:

- student potrafi samodzielnie wykorzystywać podstawowe operatory grupujące oraz funkcje agregujące i grupujące SQL
- student potrafi samodzielnie zaprojektować i zaimplementować podstawowy proces ETL
- student potrafi zaprojektować i zaimplementować prostą hurtownię danych i wykorzystać ją do przygotowania prostych raportów i wizualizacji danych
- student potrafi sformułować i wykonać podstawowe zapytania MDX

## **Wykład**

1	Zajęcia organizacyjne. Wprowadzenie do zagadnień Business Intelligence.	2
2	Operatory grupujące SQL. Funkcje agregujące i grupujące SQL	2
3	Transakcyjne a analityczne potrzeby, procesy i źródła danych	2
4	Wielowymiarowy model danych - warstwa logiczna	2
5	Podstawy hurtowni danych	2
6	Podstawy procesu ETL	2
7	Logiczna organizacja hurtowni danych	2
8	Architektura hurtowni danych	2
9	Podstawy MDX	2
10	Podstawy MDX	2
11	Wielowymiarowy model danych - warstwa fizyczna	2
12	Podstawy raportowania	2
13	Podstawy wizualizacji danych	2
14	Podstawy projektowania hurtowni danych	2
15	Webowe panele zarządzania – dashboards	2



*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

## **STOSOWANE NARZĘDZIA DYDAKTYCZNE**

- Wykład z wykorzystaniem prezentacji slajdów
- Konsultacje
- Zapoznanie się studenta z literaturą podstawową i rozszerzoną
- Ćwiczenia laboratoryjne w laboratorium komputerowym
- Praca własna studenta - przygotowanie do zajęć laboratoryjnych
- Opracowanie sprawozdania z zajęć laboratoryjnych w formie cyfrowej

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

## LITERATURA PODSTAWOWA

- Jensen C.S., Pedersen T.B., Thomsen C., Multidimensional Databases and DataWarehousing, Morgan & Claypool Publishers series SYNTHESIS LECTURES ON DATA MANAGEMENT, 2010
- Rainardi V., Building a Data Warehouse With Examples in SQL Server, Apress, 2008
- Harinath S., Pihlgren R., Lee D.G.-Y., Sirmon J., Bruckner R.M., PROFESSIONAL MICROSOFT® SQL SERVER® 2012 ANALYSIS SERVICES WITH MDX AND DAX, John Wiley & Sons, Inc., 2012
- Microsoft SQL Server 2016 Integration Services, APN Promise, 2016
- Inmon W., Building the Data Warehouse, John Wiley & Sons, New York 2002
- Kimball R., Caserta J., The Data Warehouse ETL Toolkit, Wiley Publishing, Inc, 2004

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# LITERATURA UZUPEŁNIAJĄCA

- Aspin A., SQL Server 2012 Data Integration Recipes, Apress, 2012
- Leonard A., Masson M., Mitchell T., Moss J.M., Ufford M., SQL Server 2012 Integration Services Design Patterns, Apress, 2012
- Claudia Imhoff, Nicholas Galembo, Jonathan G. Geiger, Mastering Data Warehouse Design, Wiley Publishing, Inc., 2003
- MacLennan J., Tang ZH., Crivat B., Data Mining with SQL Server 2008, Wiley Publishing, Inc, 2009

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Niezbędne oprogramowanie oraz zasoby

- MS SQL Server (Enterprise Edition)
  - Analysis Services (SSAS)
  - SQL Server Data Tools (SSIS)
  - Data Quality Services (DQS)
- Bazy danych: AdventureWorks2017, AdventureWorksDW2017
- Arkusz kalkulacyjny Excel
- Power BI
- Tableau

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Zasady zaliczenia

## Wykład

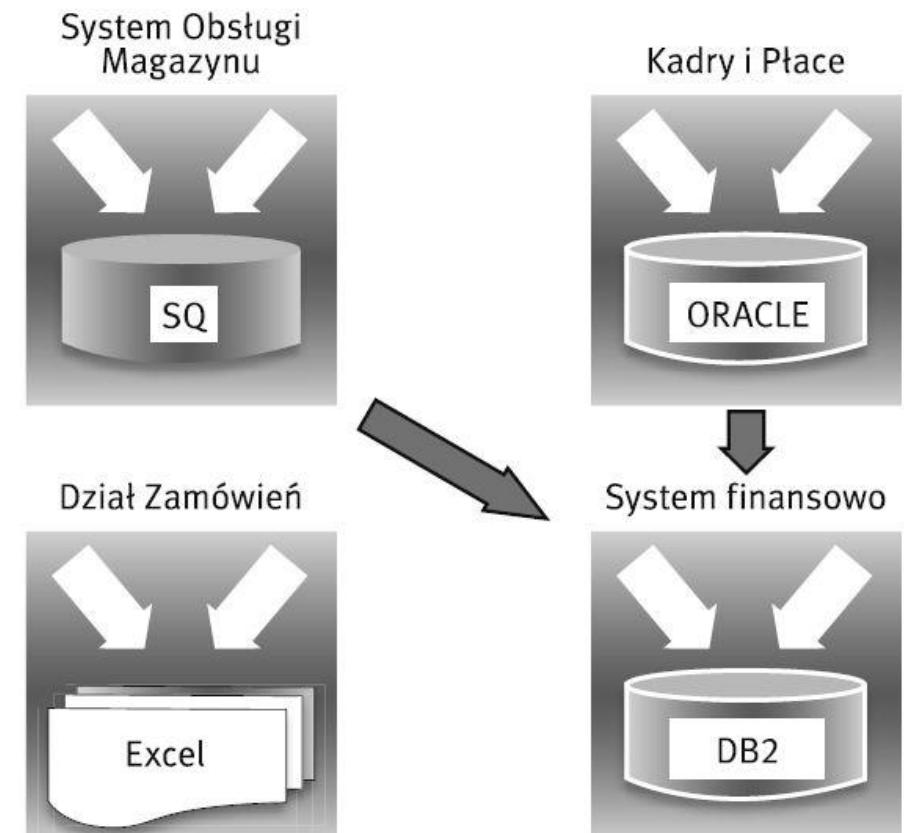
- Egzamin
  - Pierwszy termin w formie testu (eportal)
  - Drugi termin w formie testu eportal lub ustnie (Zoom)

## Laboratorium

- Warunki ustalone przez prowadzącego laboratorium

## Historia baz danych

- Różne systemy, różne aspekty działania:
  - wystawienia faktur,
  - obsługa magazynu,
  - systemy kadrowe,
  - systemy księgowie,
  - obsługa klientów
  - ...





Fundusze  
Europejskie  
Wiedza Edukacja Rozwój



Politechnika Wrocławskiego

Unia Europejska  
Europejski Fundusz Społeczny



*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# OLTP

- Przechowywane dane są zorientowane procesowo (np. wystawianie faktury)
- Stosunkowo niewielkie rozmiary baz danych (kilka gigabajtów)
- Przechowywane są dane bieżące bez konieczności gromadzenia danych historycznych
- Realizowana jest duża ilość prostych zapytań
- Przechowywane są dane elementarne
- Realizowane są operacje wstawiania, modyfikowania i usuwania danych

## OLTP

- Przechowywane dane są zorientowane procesowo (np. wystawianie faktury)
- Stosunkowo niewielkie rozmiary baz danych (kilka gigabajtów)
- Przechowywane są dane bieżące bez konieczności gromadzenia danych historycznych
- Realizowana jest duża ilość prostych zapytań
- Przechowywane są dane elementarne
- Realizowane są operacje wstawiania, modyfikowania i usuwania danych





*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

## OLTP

- Przechowywane dane są zorientowane procesowo (np. wystawianie faktury)
- Stosunkowo niewielkie rozmiary baz danych (kilka gigabajtów)
- Przechowywane są dane bieżące bez konieczności gromadzenia danych historycznych
- Realizowana jest duża ilość prostych zapytań
- Przechowywane są dane elementarne
- Realizowane są operacje wstawiania, modyfikowania i usuwania danych

*Pozyskiwanie informacji potrzebnych kierownictwu do podejmowania decyzji?*

## OLTP

- Przechowywane dane są zorientowane procesowo (np. wystawianie faktury)
- Stosunkowo niewielkie rozmiary baz danych (kilka gigabajtów)
- Przechowywane są dane bieżące bez konieczności gromadzenia danych historycznych
- Realizowana jest duża ilość prostych zapytań
- Przechowywane są dane elementarne
- Realizowane są operacje wstawiania, modyfikowania i usuwania danych

NIE WYSTARCZA!!!





## OLTP

- Przechowywane dane są zorientowane procesowo (np. wystawianie faktury)
- Stosunkowo niewielkie rozmiary baz danych (kilka gigabajtów)
- Przechowywane są dane bieżące bez konieczności gromadzenia danych historycznych
- Realizowana jest duża ilość prostych zapytań
- Przechowywane są dane elementarne
- Realizowane są operacje wstawiania, modyfikowania i usuwania danych

## OLAP

- Przechowywane dane są zorientowane tematyczne (np. sprzedaż produktów, stany zapasów)
- Bardzo duże ilości gromadzonych danych (wiele terabajtów)
- Przechowywane są dane bieżące i historyczne
- Bardzo złożone zapytania operujące na wielkich ilościach danych
- Przechowywane są dane elementarne i zagregowane (sumy, średnie)
- Wykonywane są głównie operacje dopisywania nowych danych – praktycznie nie wykonuje się operacji modyfikowania danych

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

## Hurtownia danych - definicja

**Hurtownia danych to:**

- **tematycznie zorientowana**
- **zintegrowana**
- **chronologiczna**
- **trwała**

kolekcja danych do wspomagania procesów podejmowania decyzji



Fundusze  
Europejskie  
Wiedza Edukacja Rozwój



Politechnika Wrocławska

Unia Europejska  
Europejski Fundusz Społeczny

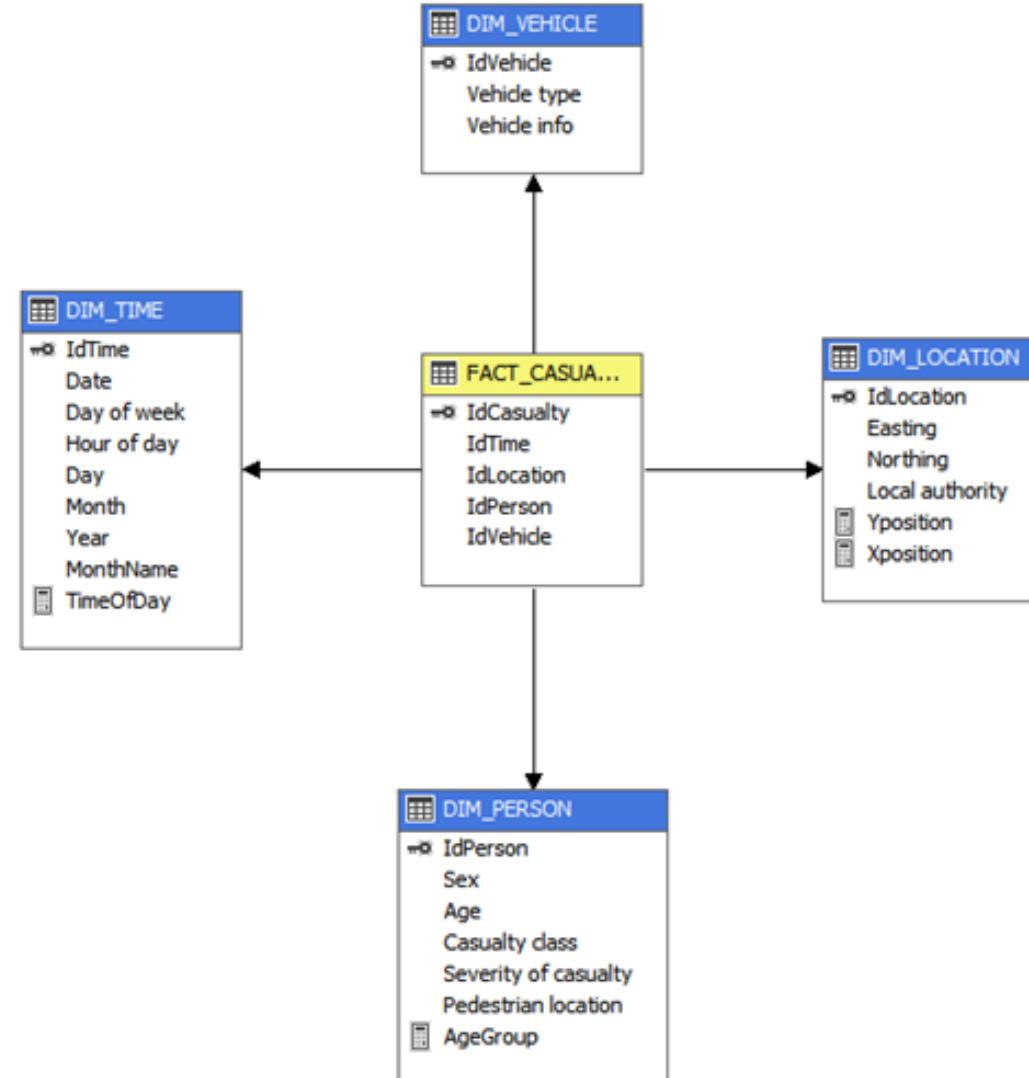


*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Przykłady hurtowni

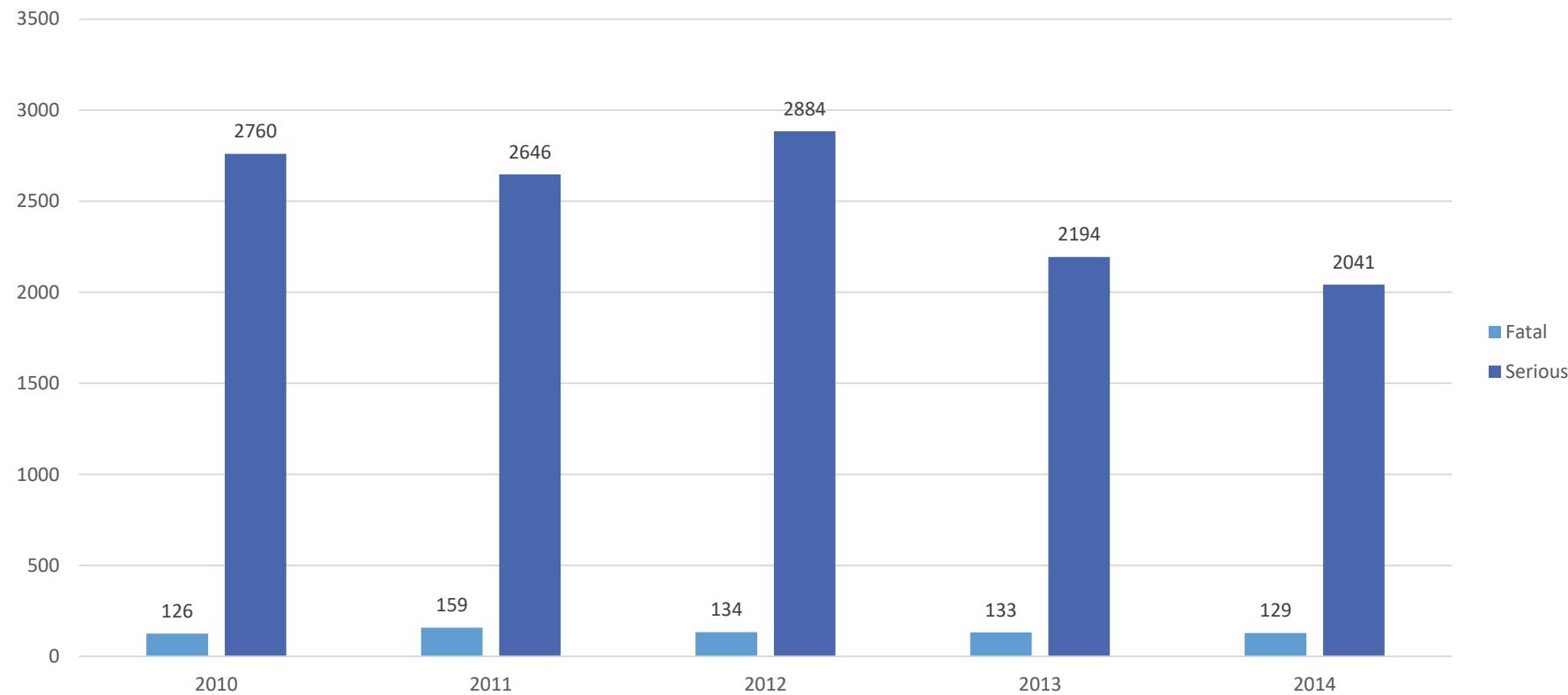


*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*



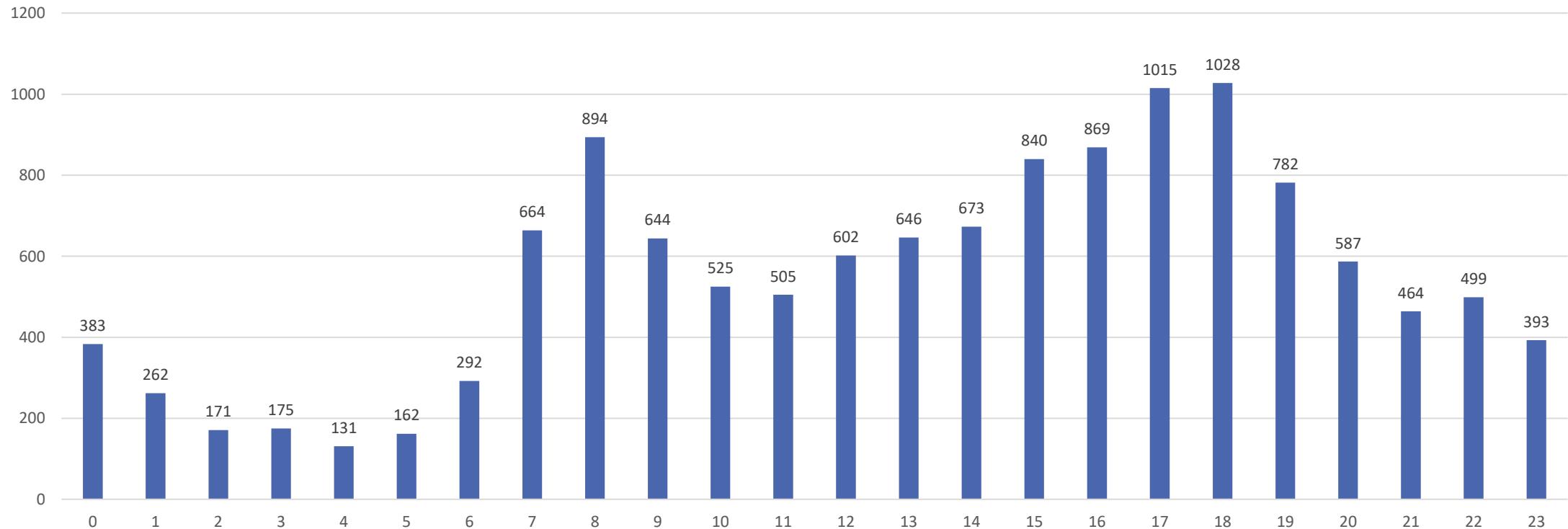
*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

## Liczba ofiar na przestrzeni lat

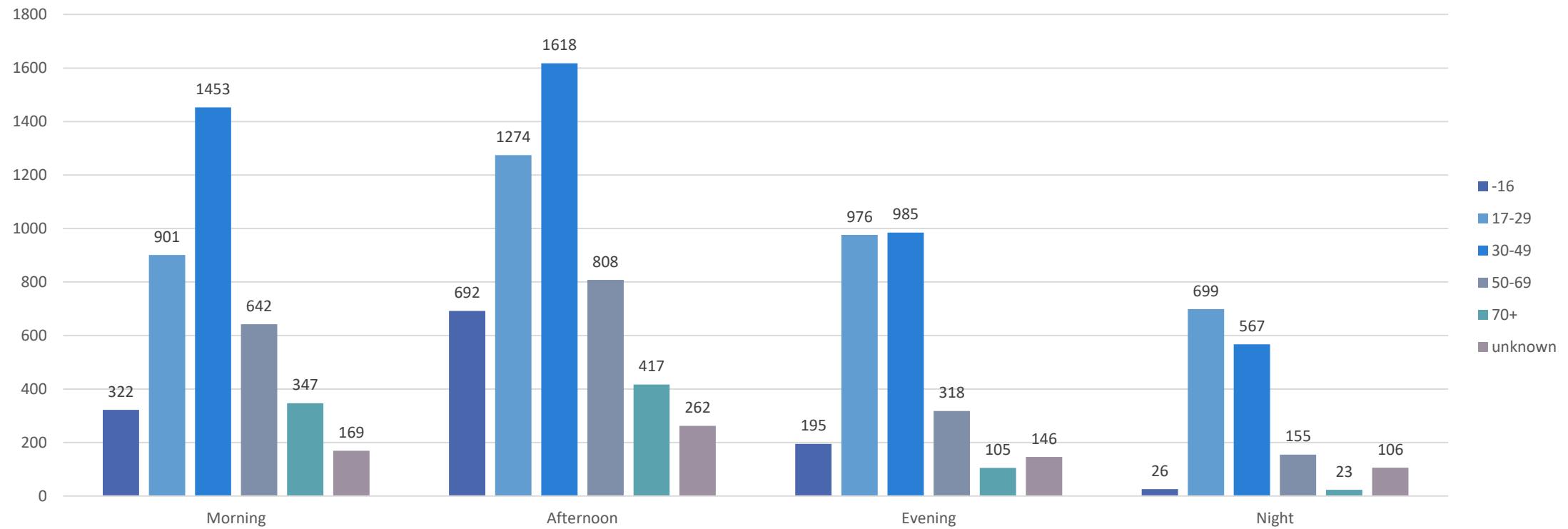


*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

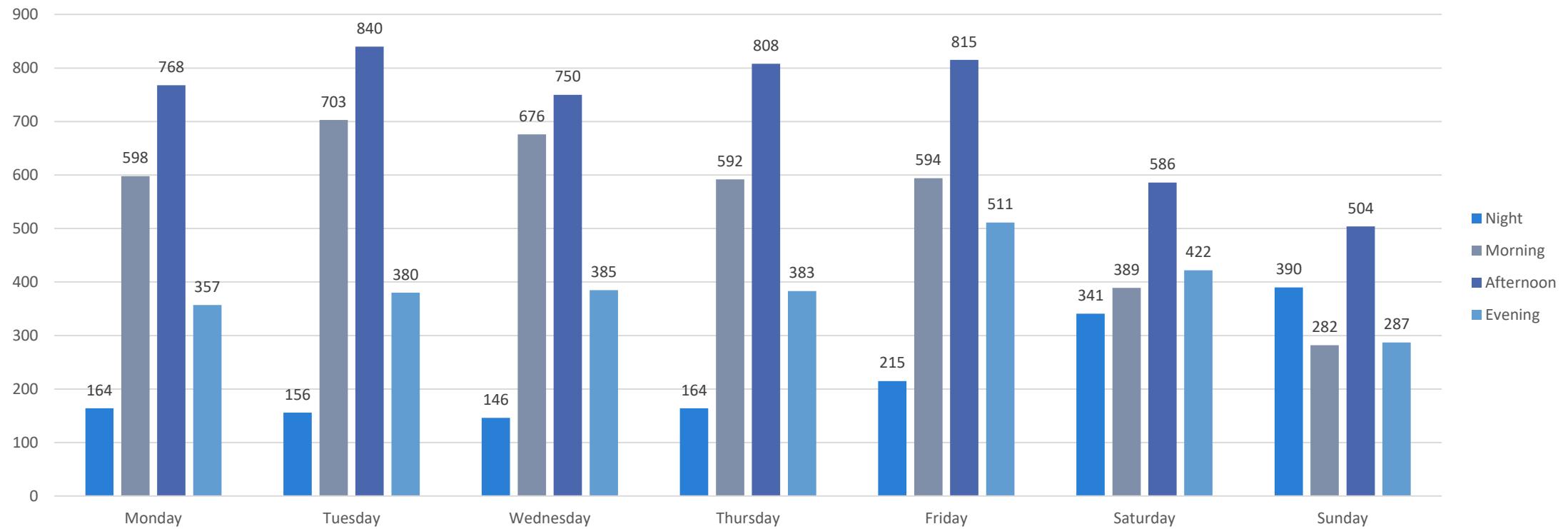
# Liczba ofiar w zależności od godziny



# Wiek ofiar w zależności od pory dnia

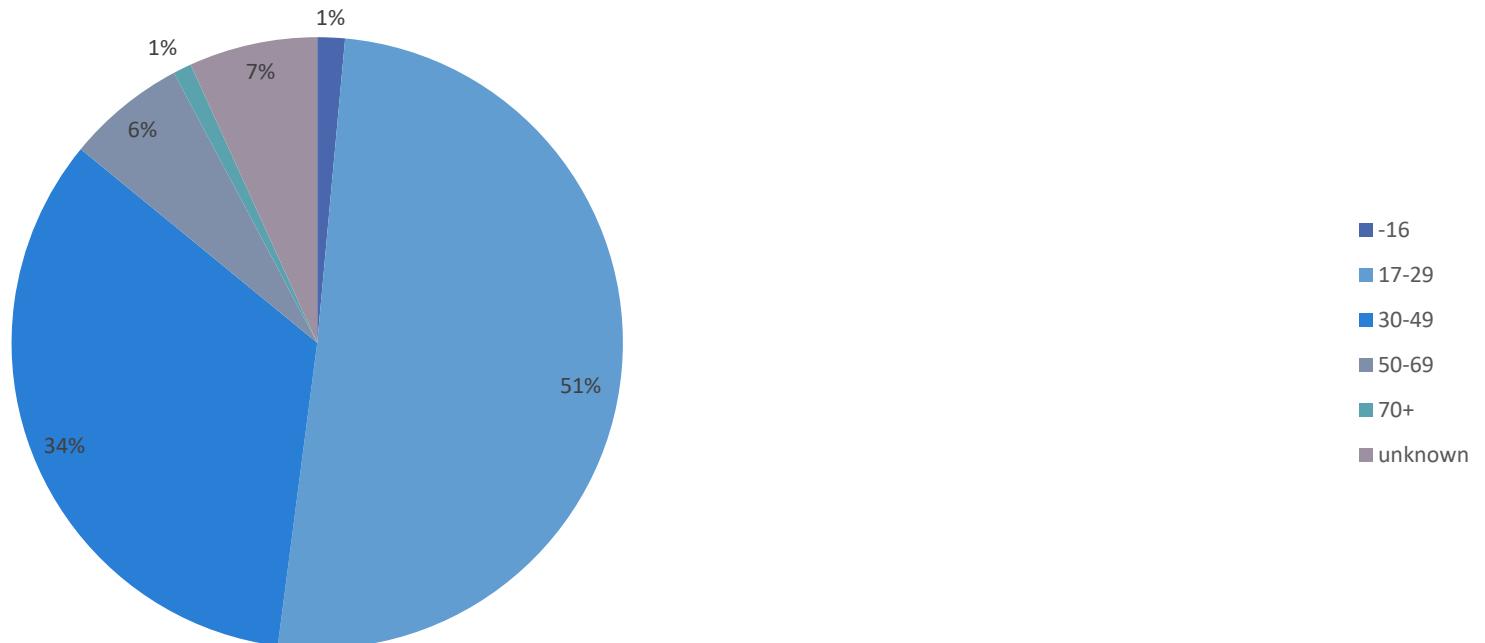


# Liczba ofiar w zależności od pory dnia w ciągu tygodnia



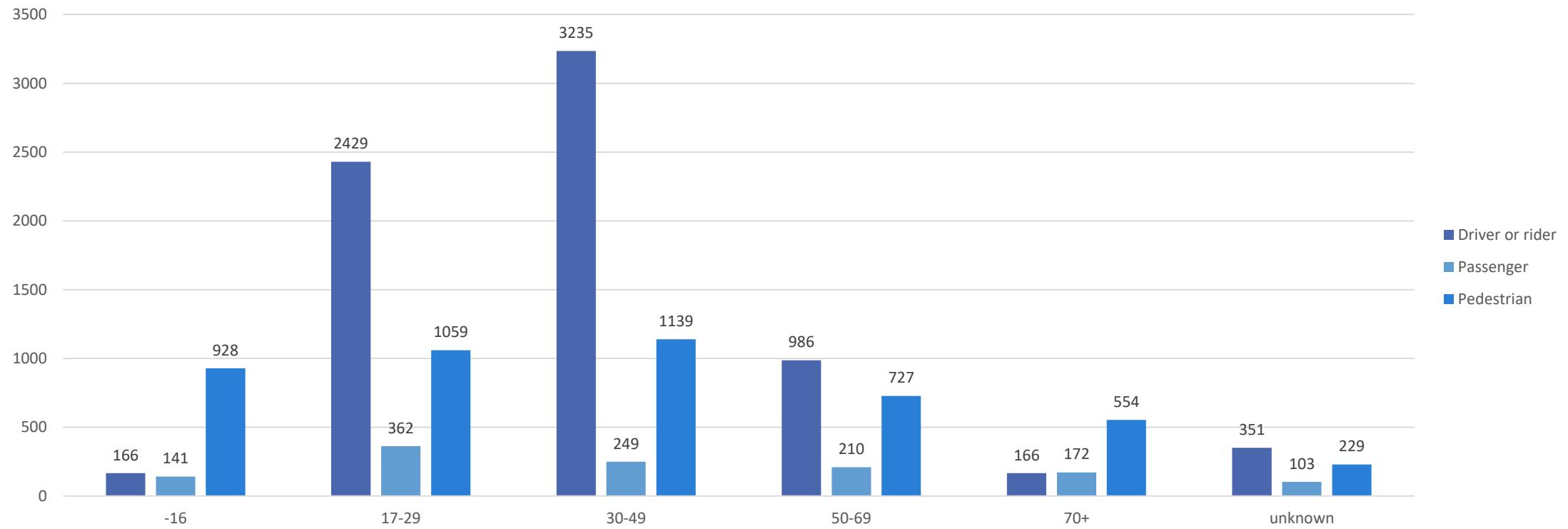
„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”

## Procent ofiar w zależności od wieku w weekendowe noce



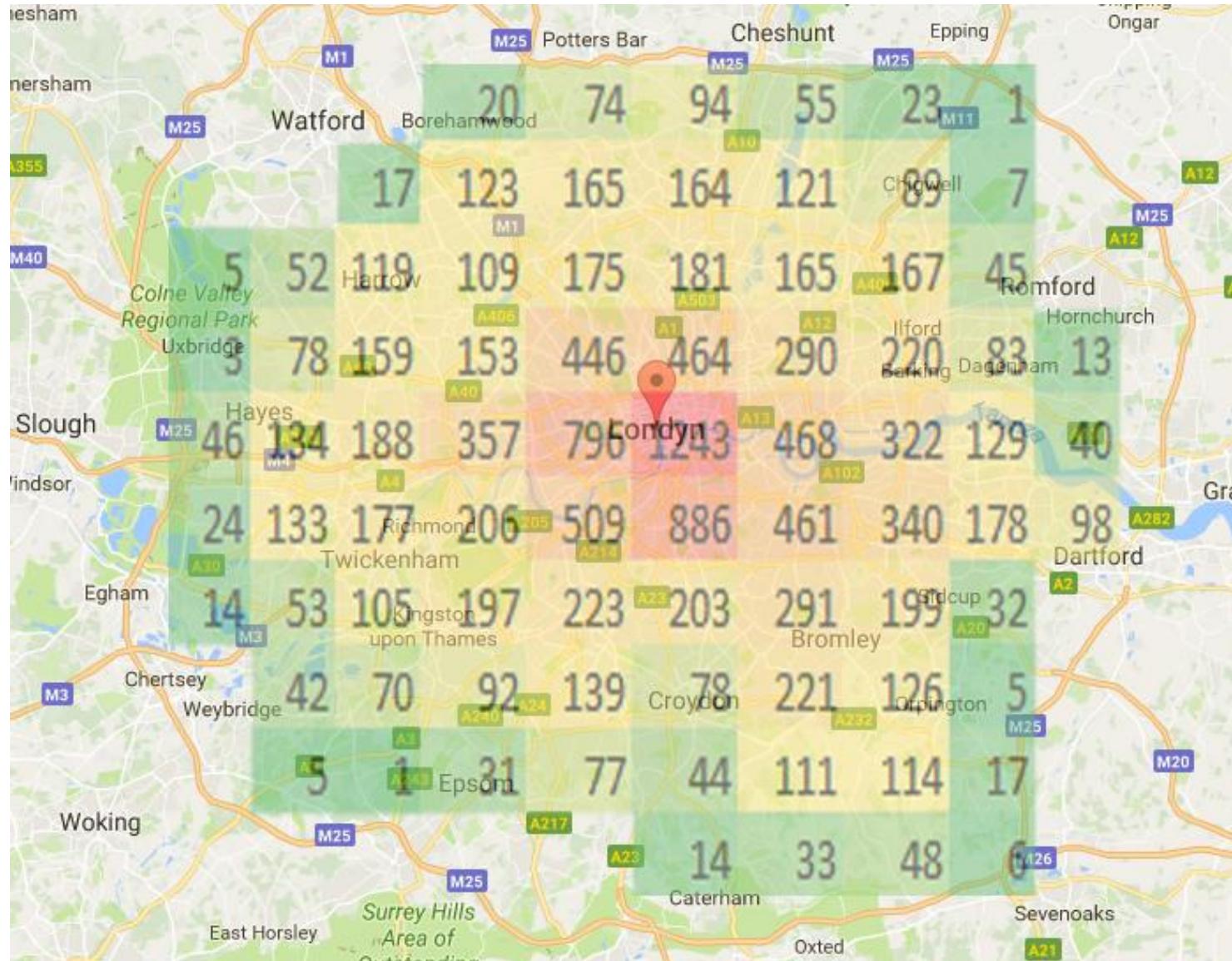
„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”

## Rodzaj ofiary w zależności od wieku



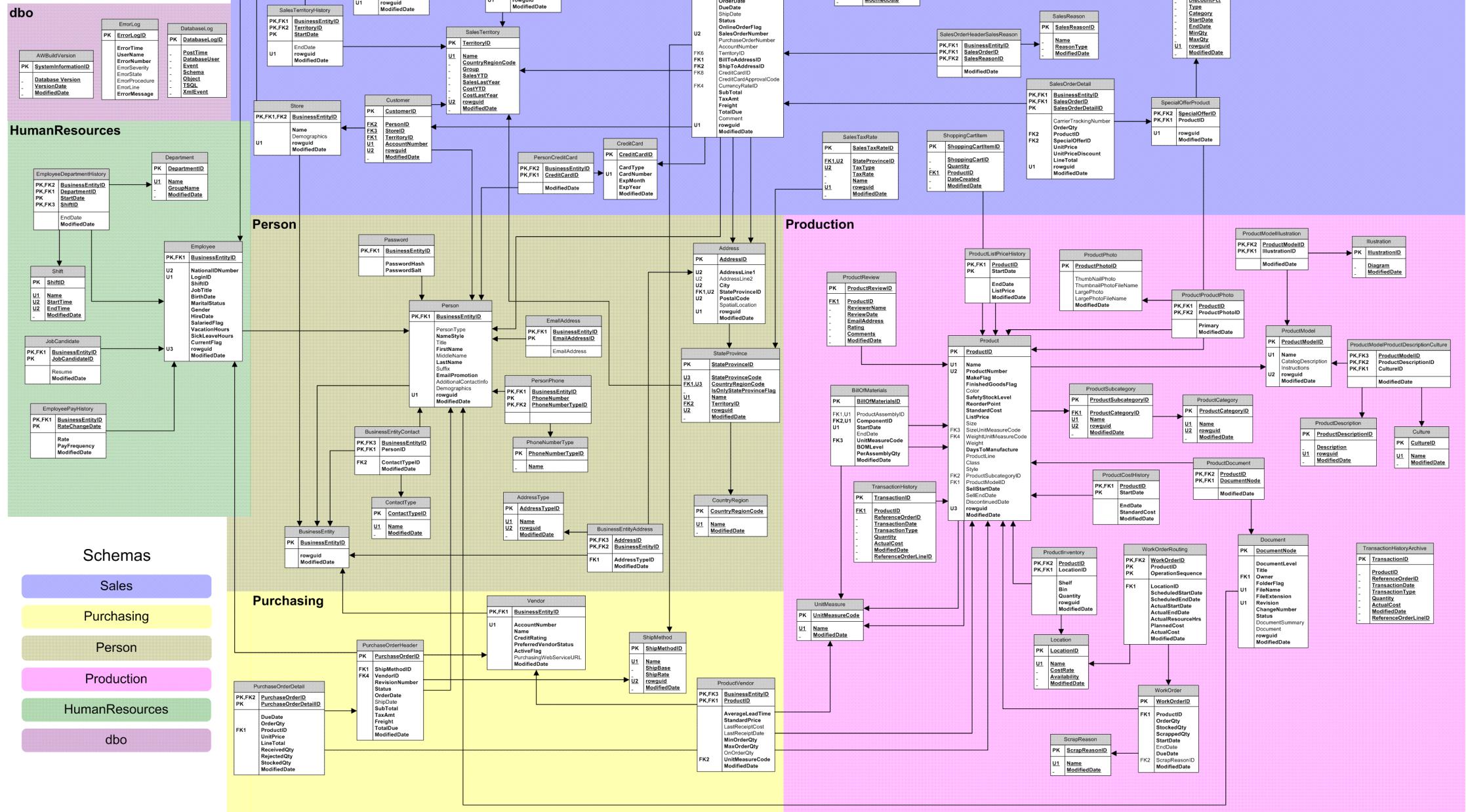


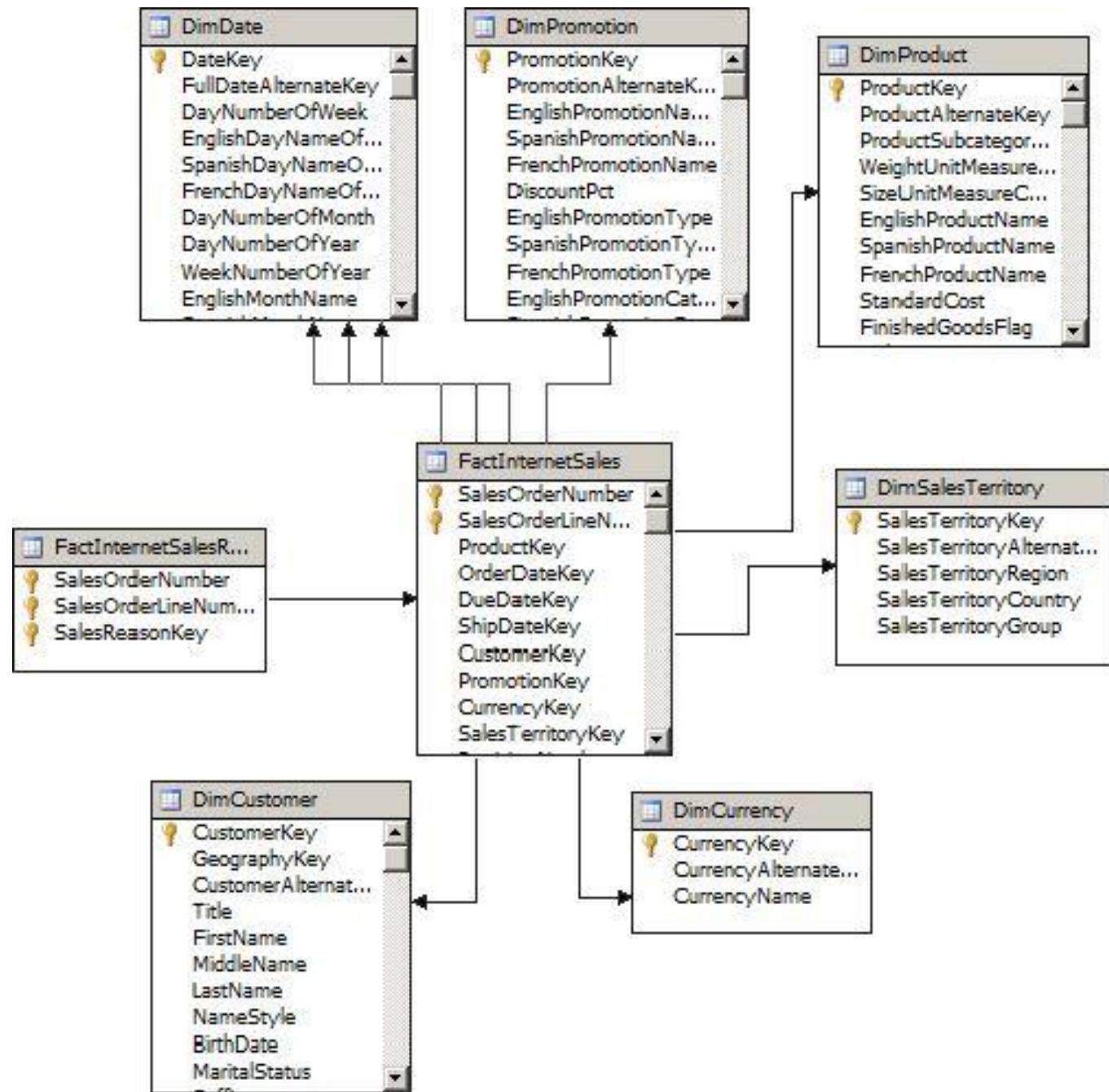
*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*



## AdventureWorks 2008 OLTP Schema

**Best Print Results if:**  
11X17 paper              Samples and Sample Databases at  
Landscape                <http://CodePlex.com/SqlServerSamples>  
Fit to 1 sheet







Fundusze  
Europejskie  
Wiedza Edukacja Rozwój



Politechnika Wrocławska

Unia Europejska  
Europejski Fundusz Społeczny



*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Hurtownie danych

Dziękuję za uwagę

dr inż. Marcin Maleszka



Fundusze  
Europejskie  
Wiedza Edukacja Rozwój



Politechnika Wrocławskiego

Unia Europejska  
Europejski Fundusz Społeczny



*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Hurtownie danych

**Operatory grupujące SQL. Funkcje agregujące i grupujące SQL**

**dr inż. Marcin Maleszka**



Fundusze  
Europejskie  
Wiedza Edukacja Rozwój



Politechnika Wrocławska

Unia Europejska  
Europejski Fundusz Społeczny



*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# SQL - podstawy

SELECT

FROM

WHERE

GROUP BY

HAVING

ORDER BY

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

## SQL - podstawy

```
SELECT <nazwy kolumn>
FROM <tabela>
WHERE <warunki>
GROUP BY <atrybuty>
HAVING <warunki dot. atrybutów zagregowanych>
ORDER BY <atrybuty>
```

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# SQL – kolejność wykonywania poleceń

FROM

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# SQL – kolejność wykonywania poleceń

FROM

WHERE

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# SQL – kolejność wykonywania poleceń

FROM

WHERE

GROUP BY

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# SQL – kolejność wykonywania poleceń

FROM

WHERE

GROUP BY

HAVING

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# SQL – kolejność wykonywania poleceń

SELECT

FROM

WHERE

GROUP BY

HAVING

# SQL – kolejność wykonywania poleceń

SELECT

FROM

WHERE

GROUP BY

HAVING

ORDER BY



„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”

## SQL - przykład

```
SELECT SalesPersonID,  
       COUNT(DISTINCT CustomerID) "Liczba klientów",  
       COUNT(*) AS "Liczba zamówień"  
FROM Sales.SalesOrderHeader  
WHERE Year(OrderDate)=2012  
GROUP BY SalesPersonID, Year(OrderDate)  
HAVING COUNT(*)>10  
ORDER BY 1,2
```

SalesOrderHeader (Sales)	
	SalesOrderID
	RevisionNumber
	OrderDate
	DueDate
	ShipDate
	Status
	OnlineOrderFlag
	SalesOrderNumber
	PurchaseOrderNumber
	AccountNumber
	CustomerID
	SalesPersonID
	TerritoryID
	BillToAddressID
	ShipToAddressID
	ShipMethodID
	CreditCardID
	CreditCardApprovalCode
	CurrencyRateID
	SubTotal
	TaxAmt
	Freight
	TotalDue
	Comment
	rowguid
	ModifiedDate



„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”

	SalesPersonID	Liczba klientów	Liczba zamówień
1	NULL	2743	2743
2	274	22	22
3	275	73	148
4	276	54	151
5	277	87	166
6	278	28	80
7	279	57	153
8	280	26	45
9	281	25	74
10	282	47	86
11	283	24	63
12	284	16	24
13	289	48	111
14	290	18	42

# Funkcje i operatory

- CASE, COALESCE, NULLIF
- CAST
- CTE
- PIVOT
- Zapytania z podsumowaniami – operatory:
  - ROLLUP
  - CUBE
  - GROUPING SETS, GROUPING, GROUPING\_ID
- Funkcje okienkowe
  - OVER PARTITION
  - Tworzenie rankingów

## Dane transakcyjne vs analityczne

	OrderID	OrderDate	OrderAmmount	CustomerName
1	1	2012-03-01	10.0000	Joe
2	2	2012-03-01	11.0000	Sam
3	3	2012-03-02	10.0000	Beth
4	4	2012-03-02	15.0000	Joe
5	5	2012-03-02	17.0000	Sam
6	6	2012-03-03	12.0000	Joe
7	7	2012-03-04	10.0000	Beth
8	8	2012-03-04	18.0000	Sam
9	9	2012-03-04	12.0000	Joe
10	10	2012-03-04	11.0000	Beth
11	11	2012-03-05	14.0000	Sam
12	12	2012-03-06	17.0000	Beth
13	13	2012-03-06	19.0000	Joe
14	14	2012-03-07	13.0000	Beth
15	15	2012-03-07	16.0000	Sam

	Sprzedawca	2011	2012	2013	2014
1	NULL	3863120,2134	6390599,9473	10732127,33	8372829,73
2	274	28926,2465	453524,5233	431088,7238	178584,3625
3	275	875823,8318	3375456,8947	3985374,8995	1057247,3786
4	276	1149715,3253	3834908,674	4111294,9056	1271088,5216
5	277	1311627,2918	4317306,5741	3396776,2674	1040093,4071
6	278	500091,8202	1283569,6294	1389836,8101	435948,9551
7	279	1521289,1881	2674436,3518	2188082,7813	787204,4289
8	280	648485,5862	1208264,3834	963420,5805	504932,044
9	281	967597,2899	2294210,5506	2387256,0616	777941,6519



Fundusze  
Europejskie  
Wiedza Edukacja Rozwój



Politechnika Wrocławskiego

Unia Europejska  
Europejski Fundusz Społeczny



„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”

## CASE

### CASE

```
WHEN plec = 0 THEN 'Kobieta'  
WHEN plec = 1 THEN 'Mężczyzna'  
WHEN plec = 2 THEN 'Nieznana'  
ELSE NULL  
END;
```

# CASE

```
SELECT CustomerID,  
       SUM(CASE WHEN DATEPART(quarter,OrderDate)=1 THEN TotalDue) AS Q1,  
       SUM(CASE WHEN DATEPART(quarter,OrderDate)=2 THEN TotalDue) AS Q2,  
       SUM(CASE WHEN DATEPART(quarter,OrderDate)=3 THEN TotalDue) AS Q3,  
       SUM(CASE WHEN DATEPART(quarter,OrderDate)=4 THEN TotalDue) AS Q4  
FROM Sales.SalesOrderHeader  
GROUP BY CustomerID
```

	CustomerID	Q1	Q2	Q3	Q4
1	14324	5659.1783	NULL	NULL	NULL
2	22814	NULL	NULL	5.514	NULL
3	11407	NULL	NULL	NULL	59.659
4	28387	NULL	645.2869	NULL	NULL
5	19897	NULL	659.6408	NULL	NULL
6	15675	2580.1529	NULL	2699.9018	2682.9953
7	24165	666.8565	2699.9018	NULL	NULL
8	27036	NULL	NULL	8.0444	NULL
9	18546	NULL	NULL	NULL	32.5754
10	11453	3729.364	NULL	2633.1377	2673.0613
11	17195	2574.9042	NULL	NULL	1105.4834
12	17026	NULL	NULL	288.836	NULL
13	22768	NULL	NULL	NULL	663.5083

## CASE - uwagi

1. Na podstawie analizy klauzuli THEN ustalany jest typ wyniku wyrażenia
2. Typ wartości musi być możliwy do wyznaczenia
3. Wynikiem jest wartość wyrażenia pierwszego spełnionego warunku WHEN
4. Można zagnieździć wyrażenie CASE
5. Jeżeli nie została jawnie określona klauzula ELSE wówczas domyślną postacią jest ELSE NULL



*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

## CASE - zadanie

- Studenci (StudentID, Imię, Nazwisko, Kierunek, Plec, Wiek)
- Ile studentów płci męskiej i żeńskiej studiuje poszczególne kierunki

SELECT Kierunek,

    SUM(CASE WHEN Plec = 'M' THEN 1 ELSE 0) AS „Mężczyźni”

    SUM(CASE WHEN Plec = 'K' THEN 1 ELSE 0) AS „Kobiety”

FROM Studenci

GROUP BY Kierunek;



Fundusze  
Europejskie  
Wiedza Edukacja Rozwój



Politechnika Wrocławska

Unia Europejska  
Europejski Fundusz Społeczny



*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# COALESCE

COALESCE (<wyrażenie 1.>, <wyrażenie 2.>)

CASE

WHEN <wyrażenie 1.> IS NOT NULL

    THEN <wyrażenie 1.>

    ELSE <wyrażenie 2.>

END;

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# NULLIF

NULLIF (<wyrażenie 1.>, <wyrażenie 2.>)

CASE

WHEN <wyrażenie 1.> = <wyrażenie 2.>

THEN NULL

ELSE <wyrażenie 1.>

END;

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Konwersja danych: CAST

- `CAST(<operand> AS <typ danych>)`
- `SELECT CAST('20190306' AS Date) "Dzisiejsze zajęcia";`
- `SELECT CAST(SYSDATETIME() AS Date) "Dzisiejsze zajęcia";`



Fundusze  
Europejskie  
Wiedza Edukacja Rozwój



Politechnika Wrocławskiego



Unia Europejska  
Europejski Fundusz Społeczny

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# CTE

- Common table expression
- nazwana tabela wirtualna zdefiniowana za pomocą wyrażenia zapytania

**WITH <nazwa CTE> [<lista kolumn>]**

**AS**

(

<definicja zapytania>

)

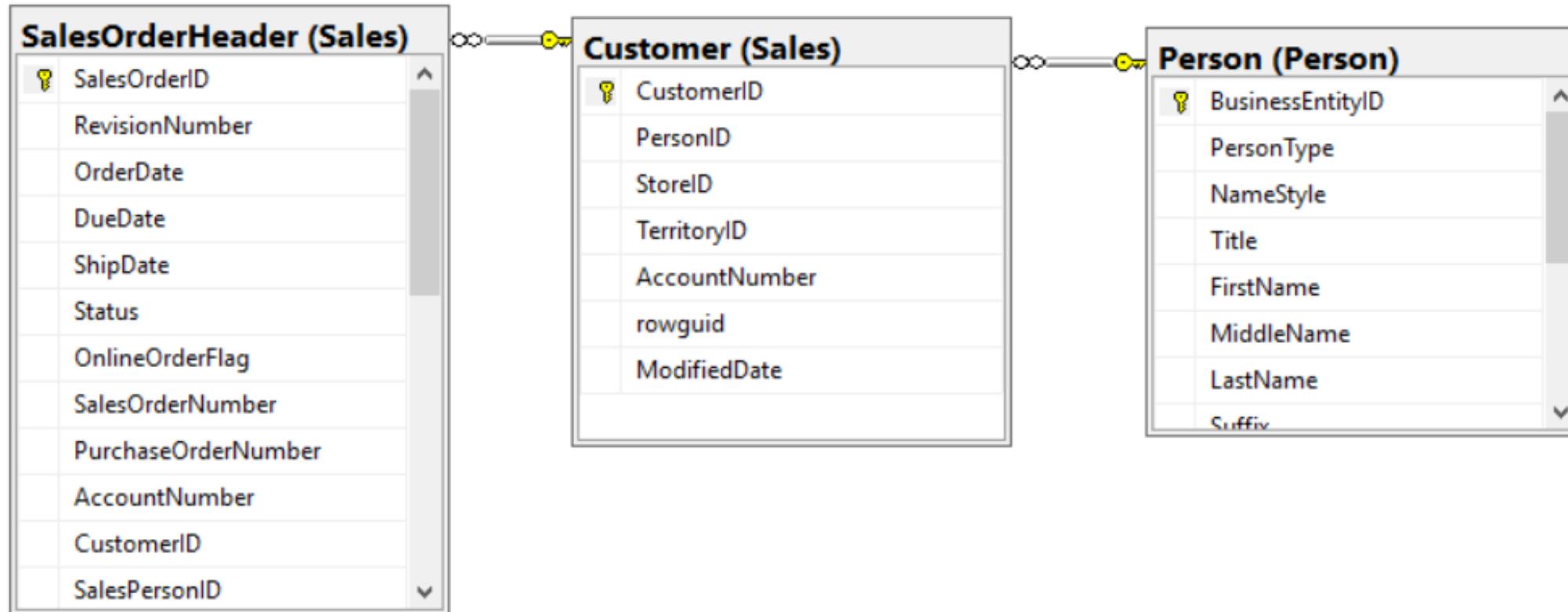
**SELECT <lista kolumn>**

**FROM <nazwa CTE>**



„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”

## CTE - przykład



SalesOrderHeader (Sales)
SalesOrderID
RevisionNumber
OrderDate
DueDate
ShipDate
Status
OnlineOrderFlag
SalesOrderNumber
PurchaseOrderNumber
AccountNumber
CustomerID
SalesPersonID

Customer (Sales)
CustomerID
PersonID
StoreID
TerritoryID
AccountNumber
rowguid
ModifiedDate

Person (Person)
BusinessEntityID
PersonType
NameStyle
Title
FirstName
MiddleName
LastName
Suffix

wska



"roju Politechniki Wrocławskiej"

## CTE - przykład

```
SELECT S.CustomerID AS "klientID",
       CONCAT(P.LastName, ', ', P.FirstName) AS
    "Nazwisko, Imię",
       COUNT(*) AS "liczba zamówień"
  FROM Sales.SalesOrderHeader s JOIN Sales.Customer c
        ON s.CustomerID = c.CustomerID
   JOIN Person.Person p
        ON p.BusinessEntityID = c.PersonID
 GROUP BY s.CustomerID
 HAVING COUNT(*) > 25;
```



„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”

## CTE - przykład

```
with liczbaZam as
(
    SELECT S.CustomerID AS "KlientID", COUNT(*)
AS "liczba zamówień"
        FROM Sales.SalesOrderHeader S
        GROUP BY S.CustomerID
        HAVING COUNT(*) > 25
)
```



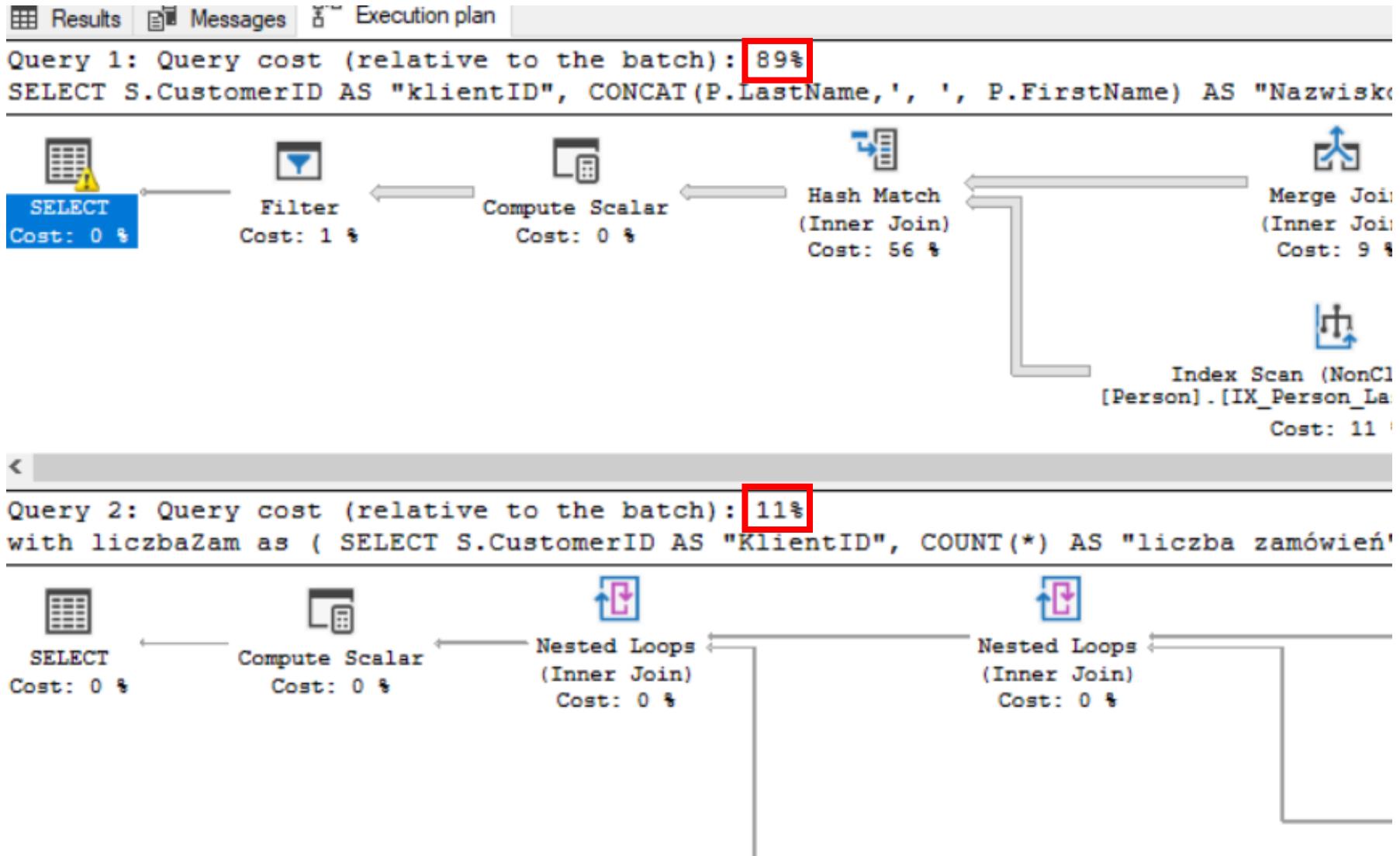
„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”

## CTE - przykład

```
SELECT klientID,
       CONCAT(P.LastName, ', ', P.FirstName) AS "Nazwisko, Imię"
     , z.[liczba zamówień]
  FROM liczbaZam z JOIN Sales.Customer C
    ON z.klientID = C.CustomerID
   JOIN Person.Person P
    ON P.BusinessEntityID = c.PersonID
;
```



„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”



# PIVOT

- Grupowanie
- Rozpraszanie
- Agregowanie

SELECT ...

FROM <źródło danych>

PIVOT (<funkcja agr.>(<element>)

FOR <element podziału> IN (<lista kolumn>)

)AS <alias tab. wynikowej>

niejawne grupowanie wierszy tabeli  
na podstawie wszystkich jej kolumn  
nie wymienionych jako argument operatora PIVOT

rozpraszanie wartości kolumny tabeli - CASE

agregacja dla każdego wyrażenia CASE



„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”

# PIVOT

```
SELECT SalesPersonID
      , CAST([2011] AS DEC(10,2)) [2011]
      , CAST([2012] AS DEC(10,2)) [2012]
      , CAST([2013] AS DEC(10,2)) [2013]
FROM (SELECT SalesPersonID
       , YEAR(OrderDate) AS Od, SubTotal
     FROM Sales.SalesOrderHeader) AS e
PIVOT (SUM(SubTotal)
      FOR Od IN ([2011],[2012], [2013])) AS X;
```

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

## PIVOT - wynik

SalesPersonID	2011	2012	2013
NULL	3863120.21	6390599.95	10732127.33
274	28926.25	453524.52	431088.72
275	875823.83	3375456.89	3985374.90
276	1149715.33	3834908.67	4111294.91
277	1311627.29	4317306.57	3396776.27
278	500091.82	1283569.63	1389836.81
279	1521289.19	2674436.35	2188082.78
280	648485.59	1208264.38	963420.58
281	967597.29	2294210.55	2387256.06



„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”

## PIVOT - zadanie

- Jaki jest wynik zapytania:

```
SELECT SalesPersonID, [29491], [29605]
FROM
  (SELECT SalesPersonID, CustomerID, SubTotal
   FROM Sales.SalesOrderHeader) S
  PIVOT (
    SUM(SubTotal) FOR CustomerID IN ([29491],
      [29605])) AS X
WHERE SalesPersonID IS NOT NULL;
```

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

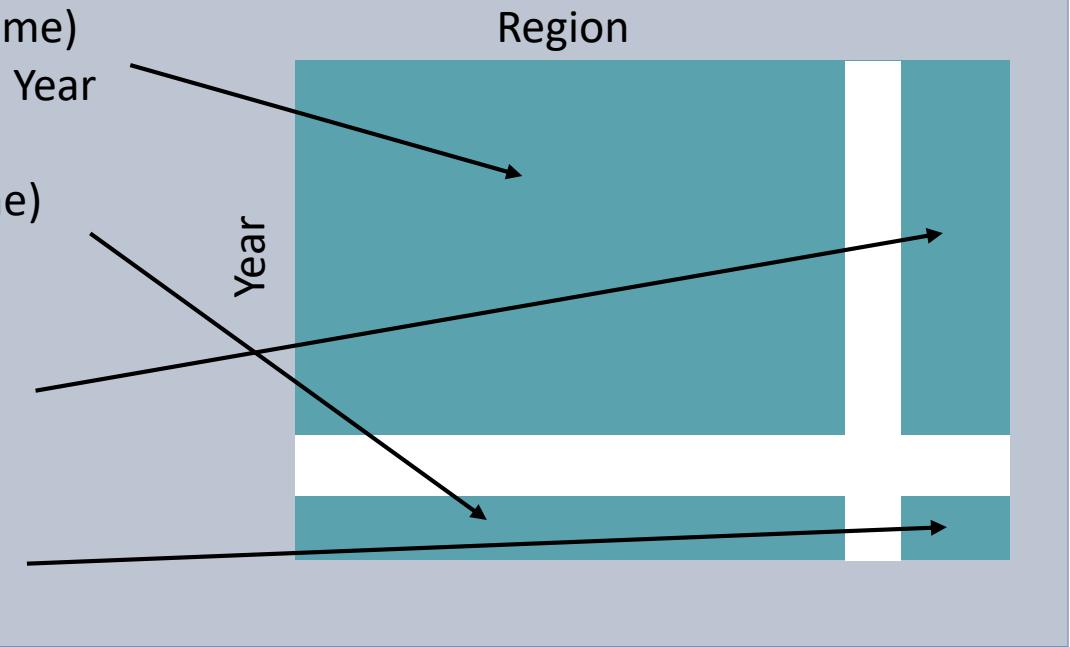
## PIVOT - zadanie

SalesPersonID	29491	29605
278	NULL	NULL
275	53863,4443	NULL
284	NULL	NULL
281	NULL	218418,7057
287	NULL	NULL
290	NULL	NULL
282	NULL	NULL
274	33406,7043	29482,0603

## Zapytania z podsumowaniami

- Sales (Region, Year, Income)
- UNION
- UNION ALL

```
SELECT Region, Year, SUM(Income)
FROM Sales GROUP BY Region, Year
UNION ALL
SELECT Region, '*', SUM(Income)
FROM Sales GROUP BY Region
UNION ALL
SELECT '*', Year, SUM(Income)
FROM Sales GROUP BY Year
UNION ALL
SELECT '*', '*', SUM(Income)
FROM Sales
```





Fundusze  
Europejskie  
Wiedza Edukacja Rozwój



Politechnika Wrocławska

Unia Europejska  
Europejski Fundusz Społeczny



*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Zapytania z podsumowaniami

- ROLLUP
- CUBE
- GROUPING SETS



Fundusze  
Europejskie  
Wiedza Edukacja Rozwój



Politechnika Wrocławska

Unia Europejska  
Europejski Fundusz Społeczny



*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

## ROLLUP

```
SELECT Region, Year, SUM(Income)
FROM Sales
GROUP BY ROLLUP(Region, Year);
```



Fundusze  
Europejskie  
Wiedza Edukacja Rozwój



Politechnika Wrocławska

Unia Europejska  
Europejski Fundusz Społeczny



*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# CUBE

```
SELECT Region, Year, SUM(Income)  
FROM Sales  
GROUP BY CUBE(Region, Year);
```



Fundusze  
Europejskie  
Wiedza Edukacja Rozwój



Politechnika Wrocławska

Unia Europejska  
Europejski Fundusz Społeczny



*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

## GROUPING SETS

```
SELECT Region, Year, SUM(Income)  
FROM Sales  
GROUP BY GROUPING SETS(Region, Year);
```

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

## Zapytania z podsumowaniami **GROUP BY .... (a, b, c)**

- ROLLUP
  - (), (a), (a, b), (a, b, c)
- CUBE
  - (), (a), (b), (c), (a, b), (a, c), (b, c), (a, b, c)
- GROUP BY GROUPING SETS ((a), (c))
  - (a), (c)

## Zapytania z podsumowaniami - przykład

```
SELECT salesPersonID, customerID, SUM(subTotal)
FROM Sales.SalesOrderHeader
GROUP BY CUBE(salesPersonID, customerID)
```

pracID	klientID	Suma
NULL	NULL	127337180.11
NULL	1	102351.80
NULL	2	29623.50
NULL	3	433942.38
NULL	29483	2049.10
268	NULL	1369624.65
268	7	3569.43
268	38	2785.51

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

## Funkcje okienkowe

- GROUP BY – dla każdej grupy jedna wartość wyrażenia
- OVER – wgląd przez „okno rekordów”

```
<funkcja agregująca> OVER (
    [<klauzula partycjonowania>]
    [<klauzula ORDER> [<klauzula ramki>] ]
)
```

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

## Funkcje okienkowe

- GROUP BY – dla każdej grupy jedna wartość wyrażenia
- OVER – wgląd przez „okno rekordów”

```
<funkcja agregująca> OVER (
    PARTITION BY <lista kolumn>
    ORDER BY <lista porządkowania>
)
```

# OVER

- Funkcje agregujące:
  - COUNT, SUM, AVG, MIN, MAX
- Funkcje szeregujące:
  - ROW\_NUMBER – nr pozycji
  - RANK – ranking (to samo miejsce dla tych samych wartości)
  - DENSE\_RANK – ranking, numerowanie ciągłe
  - NTILE – grupuje rekordy poprzez przypisanie tej samej wartości szeregującej członkom grupy



„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”

## OVER - przykład

```
SELECT DISTINCT CustomerID,  
    SUM(TotalDue) OVER(Partition BY CustomerID) "Suma zakupów",  
    AVG(TotalDue) OVER(Partition BY CustomerID) "Średnia zakupów"  
FROM Sales.SalesOrderHeader;
```

CustomerID	Suma zakupów	Średnia zakupów
15254	1795,0173	897,5086
16321	107,4613	53,7306
18727	95,3284	47,6642
24930	71,7919	71,7919
18470	6653,8902	3326,9451



„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”

## OVER - przykład

```
SELECT SalesOrderID, CustomerID,  
       RANK() OVER(ORDER BY CustomerID) AS Ranking  
FROM Sales.SalesOrderHeader;
```

```
SELECT SalesOrderID, CustomerID,  
       DENSE_RANK() OVER(ORDER BY CustomerID) AS  
           "Dense ranking"  
FROM Sales.SalesOrderHeader;
```

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# OVER - wynik

SalesOrdesID	Average	Ranking RANK	Dense Ranking DENSE_RANK
51131	151704,9021	1	1
55282	151704,9021	1	1
61324	151704,9021	1	1
65322	151704,9021	1	1
54356	123246,6745	5	2
52456	123246,6745	5	2



„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”

## OVER - przykład

```
SELECT SalesOrderID, CustomerID, SubTotal,  
       SUM(SubTotal) OVER(PARTITION BY CustomerID) AS Sprzedaż,  
       SUM(SubTotal) OVER() AS "Całkowita sprzedaż",  
       SUM(SubTotal) OVER(PARTITION BY CustomerID)  
           / SUM(SubTotal) OVER()*100.    AS "Udział klienta %"  
FROM Sales.SalesOrderHeader
```

SalesOrderID	CustomerID	SubTotal	Sprzedaż	Całkowita sprzedaż	Udział klienta %
51783	29818	107270,1188	877107,1923	109846381,4039	0.79
50270	29818	78041,5646	877107,1923	109846381,4039	0.79
57105	29818	110050,8354	877107,1923	109846381,4039	0.79
45577	29715	58704,1822	853849,1795	109846381,4039	0.77
44801	29715	48156,2081	853849,1795	109846381,4039	0.77
44133	29715	25686,5186	853849,1795	109846381,4039	0.77



„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”

## NTILE - przykład

```
WITH Sprzedaz AS
(
    SELECT [SalesPersonID] emp, AVG([SubTotal]) srednia
    FROM [Sales].[SalesOrderHeader]
    WHERE [SalesPersonID] IS NOT NULL
    GROUP BY [SalesPersonID]
)

SELECT srednia, emp
    , NTILE(5) OVER(ORDER BY srednia DESC) AS grupy
FROM Sprzedaz;
```



## NTILE - wynik

```
WITH Sprzedaz AS
(
    SELECT [SalesPersonID] emp, AVG([SubTotal])
    FROM [Sales].[SalesOrderHeader]
    WHERE [SalesPersonID] IS NOT NULL
    GROUP BY [SalesPersonID]
)
SELECT srednia, emp
    , NTILE(5) OVER(ORDER BY srednia DESC)
FROM Sprzedaz;
```

srednia	emp	grupy
35001,0799	280	1
26557,8741	281	1
25770,7938	290	1
24801,4531	276	2
24434,8811	289	2
22752,5803	274	2
21868,7024	282	3
21280,7685	277	3
20653,1177	275	3
19735,1605	283	4
18788,697	287	4

średnia	emp	Ocena pracownika
35001,0799	280	Wyróżniony
26557,8741	281	Wyróżniony
25770,7938	290	Wyróżniony
21868,7024	282	Dobry
21280,7685	277	Dobry
16518,1835	284	Kandydat do zwolnienia
15424,988	278	Kandydat do zwolnienia

```

WITH Sprzedaz AS
(
    SELECT [SalesPersonID] emp, AVG([SubTotal]) srednia
    FROM [Sales].[SalesOrderHeader]
    WHERE [SalesPersonID] IS NOT NULL
    GROUP BY [SalesPersonID]
)
SELECT srednia, emp,
    CASE NTILE(3) OVER(ORDER BY srednia DESC)
        WHEN 1 THEN 'Wyróżniony'
        WHEN 2 THEN 'Dobry'
        WHEN 3 THEN 'Kandydat do zwolnienia'
    END AS "Ocena Pracownika"
FROM Sprzedaz

```



„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”

# OVER

```
SELECT *
FROM
(SELECT DISTINCT CustomerID, -- SalesOrderID,
SUM(SubTotal) OVER(PARTITION BY CustomerID) Suma,
AVG(SubTotal) OVER(PARTITION BY CustomerID) AS
                      Srednia,
ROW_NUMBER() OVER(ORDER BY CustomerID) AS RNUM,
DENSE_RANK() OVER(ORDER BY CustomerID) AS
                      denseRANKtest,
RANK() OVER(ORDER BY CustomerID) AS RANKtest,
NTILE(1000) OVER(ORDER BY CustomerID) AS NTILEtest
FROM Sales.SalesOrderHeader) T1
```



„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”

## OVER - wynik

CustomerID	Suma	Srednia	RNUM	denseRANK test	RANKtest	NTILEtest
11000	8248,99	2749,6633	1	1	1	1
11000	8248,99	2749,6633	2	1	1	1
11000	8248,99	2749,6633	3	1	1	1
11001	6383,88	2127,96	4	2	4	1
11001	6383,88	2127,96	5	2	4	1
11001	6383,88	2127,96	6	2	4	1
11002	8114,04	2704,68	7	3	7	1
11002	8114,04	2704,68	8	3	7	1
11002	8114,04	2704,68	9	3	7	1
11003	8139,29	2713,0966	10	4	10	1

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

## Pozycjonowanie wyniku

- LAG, LEAD, FIRST\_VALUE, LAST\_VALUE
  - Znajdują wiersz w podziale i odczytują jego wartość
- OVER (PARTITION BY (Year(OrderDate)) ORDER BY MONTH(OrderDate) ROWS BETWEEN 1 PRECEDING AND CURRENT ROW
  - W bieżącym i poprzednim miesiącu
- CUME\_DIST, PERCENT\_RANK
  - Wyznacza percentile



Fundusze  
Europejskie  
Wiedza Edukacja Rozwój



Politechnika Wrocławska

Unia Europejska  
Europejski Fundusz Społeczny



*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Hurtownie danych

Dziękuję za uwagę

dr inż. Marcin Maleszka



Fundusze  
Europejskie  
Wiedza Edukacja Rozwój



Politechnika Wrocławskiego

Unia Europejska  
Europejski Fundusz Społeczny



*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Hurtownie danych

**Transakcyjne a analityczne potrzeby, procesy i źródła danych**

**dr inż. Marcin Maleszka**

## Dane -> informacje -> wiedza

- Dane, informacje – najważniejsze wartości firmy
- Cele:
  - operacyjne – gromadzenie i przetwarzanie
  - analityczne – uzyskiwanie informacji z danych i podejmowanie decyzji
    - zarządzanie
- Zarządzanie jest sztuką **zadawania właściwych pytań i podejmowania odpowiednich działań biznesowych w zależności od otrzymanych odpowiedzi**

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Dlaczego OLTP nie wystarcza?

- Operacje transakcyjne i analityczne na tej samej bazie:
  - modyfikacje
  - agregacje
- Problemy:
  - potrzeba doświadczenia i umiejętności pracownika
  - wydajność
- Rozwiązanie:
  - separacja danych

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

## Hurtownia danych - definicja

**Hurtownia danych to:**

- **tematycznie zorientowana**
- **zintegrowana**
- **chronologiczna**
- **trwała**

kolekcja danych do wspomagania procesów podejmowania decyzji

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Hurtownia danych

- **Hurtownia danych to:**
  - **tematycznie zorientowana**
  - **zintegrowana – dane** skonsolidowane (łączone, scalane, transformowane) dane pozyskiwane z różnych źródeł (systemów, arkuszy Excel, innych źródeł)
  - **chronologiczna**
  - **trwała**
- kolekcja danych do wspomagania procesów podejmowania decyzji
- Dane przechowywane w hurtowni są spójne, zintegrowane, łatwo (technicznie) dostępne i gromadzone na przestrzeni czasu (historyczne).

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

## Hurtownia danych

- „Hurtownia danych to kopia danych transakcyjnych zaprojektowana pod kątem zapytań i raportowania”

***Ralph Kimball „The Data Warehouse Toolkit”***

- „Hurtownia danych to osobne repozytorium danych, w którym informacja jest przechowywana w formacie odpowiednim dla systemów Business Intelligence i DSS”

***SAS Rapid Data Warehousing***

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Hurtownia danych - procesy

- Hurtownia danych wiąże się z procesem:
  - pozyskiwania,
  - „czyszczenia” (weryfikowania, uzgadniania),
  - przetwarzania danych pozyskiwanych z różnych źródeł do poziomu informacji, która jest zrozumiała i dostępna do odbiorcy biznesowego
- Hurtownia danych stanowi jednolitą platformę informacyjną do wszechstronnych zastosowań w dziedzinie systemów wspomagających zarządzanie

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Separacja danych

- Hurtownia – kopia danych przygotowanych do celów analitycznych
- DBMS – OLTP:
  - modyfikacja stanu bazy danych
  - INSERT, UPDATE, DELETE, MERGE
- Hurtownia danych – OLAP:
  - kompleksowe raporty
  - widoki wielowymiarowe
  - konsolidacje

# Mity i fakty

- **Mity na temat hurtowni danych (HD):**
  - HD to „bardzo duża” baza danych
  - HD wymaga specjalnych modeli przechowywania danych
  - HD to głównie wyzwania w zakresie technologii
  - Hurtownia danych = Business Intelligence
- „*Koncepcja zarządzania mająca na celu zapewnienie menedżerom informacji o odpowiedniej jakości i w odpowiednim momencie*”

„Informatyka narzędziem zarządzania w XXI wieku” red. J. Kisielnicki

- **Hurtownia danych to:**
  - Kolekcja danych=f(atrbuty, cel)
  - Model, koncepcja, filozofia zarządzania danymi w przedsiębiorstwie

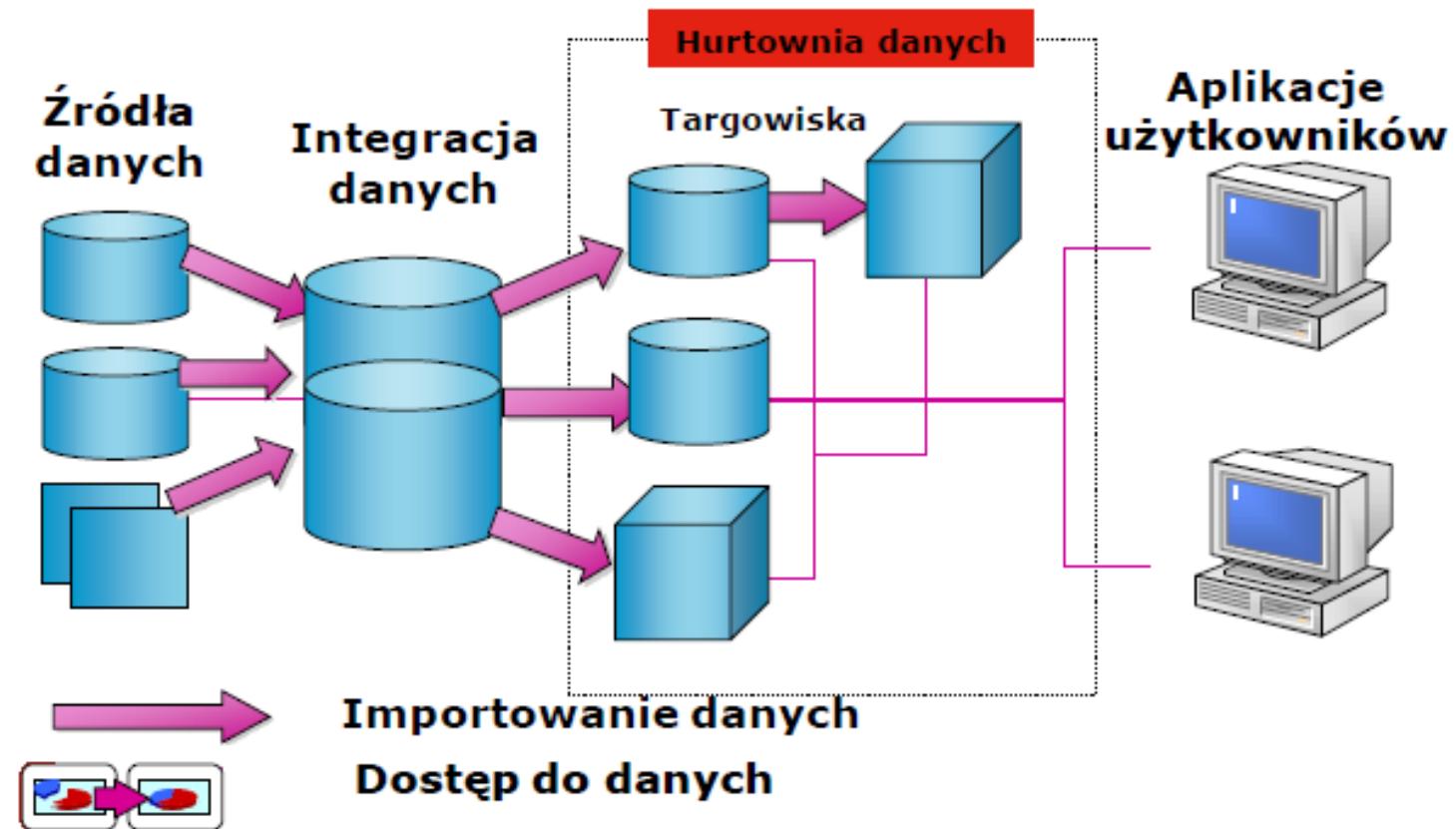


„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”

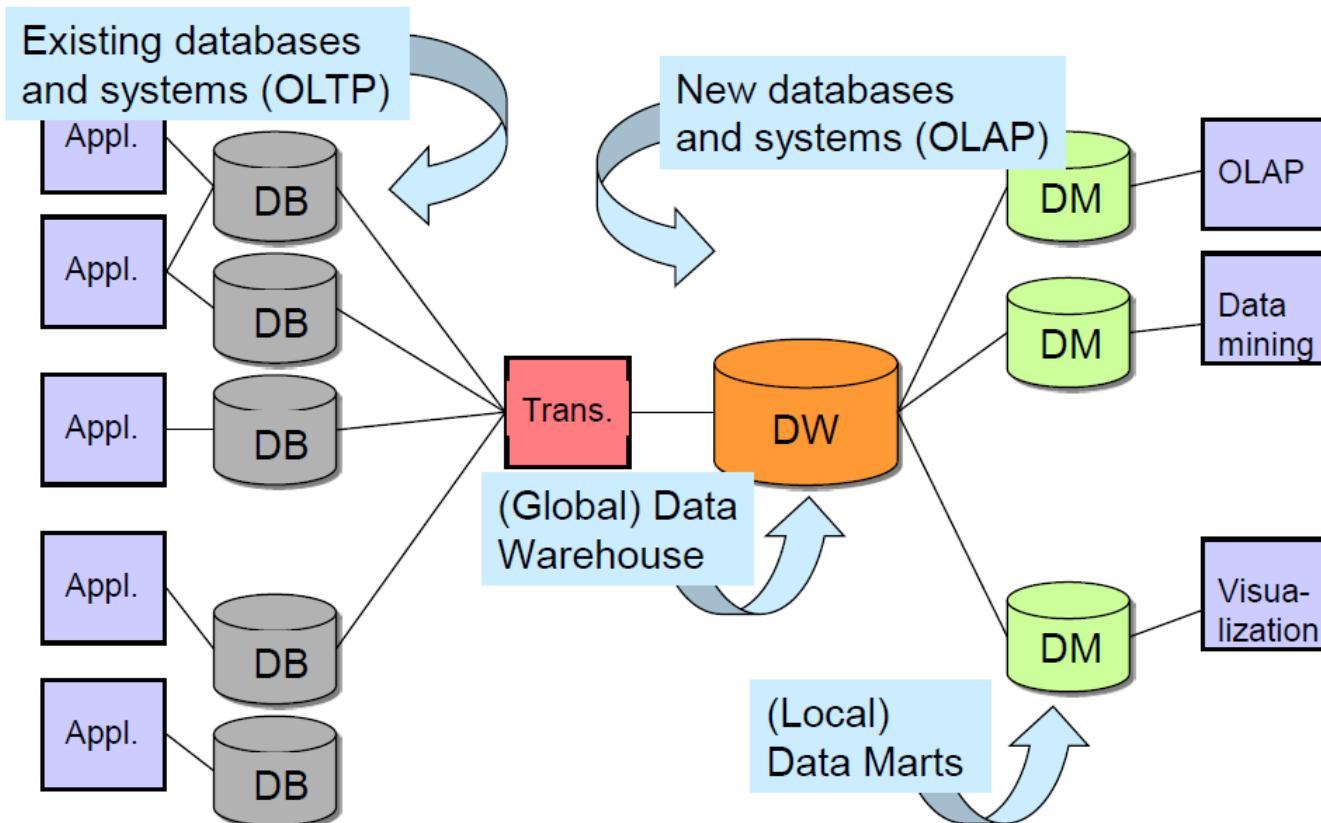
## HD vs OLTP

Aspekt	HD	OLTP
Tematycznie zorientowana	Ścisły podział na obszary informacyjne (tematyczne); lokalizacja danych zależy od tematyki	Aplikacje projektowane są wokół procesów oraz funkcji, które wymagają różnych danych
Zintegrowana	Spójność danych, kluczy, synonimy, homonimy, rekordy logiczne...	Jedno źródło, bez integracji
Chronologiczna	Dane są znakowane czasowo, szerszy horyzont czasowy np. 5 lat, historyzacja zmian	Dane są znakowane czasowo w wąskim horyzoncie np. 30-90 dni
Trwała	Dane nie są modyfikowane przez użytkowników, odczyt danych	Transakcje (insert, delete, update)
Cel	Wspomaganie procesów podejmowania decyzji (raczej na poziomach: taktycznym i strategicznym)	Ewidencja transakcji, wspieranie operacyjnego poziomu zarządzania (również podejmowania decyzji na tym poziomie)

# Hurtownia danych



# Separacja danych



Analogy: (data) producers ↔ warehouse ↔ (data) consumers

# Separacja danych

- W celu zapewnienia akceptowalnej jakości usług transakcyjnych i analitycznych dokonano odseparowania danych analitycznych od transakcyjnych
- Efekt:
  - Bazy transakcyjne (dane znormalizowane)
  - Bazy analityczne -> **Hurtownie danych**(Data Warehouse) – dane zdenormalizowane, wstępnie zagregowane

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Projektowanie hurtowni

- Opracowanie modelu:
  - pojęciowego – definicje pojęć, opis struktury, zawartości i przeznaczenia
  - logicznego – opis logiczny bazy danych i procesów hurtowni
  - fizycznego – cechy implementacyjne: indeksowanie, partycjonowanie, archiwizacje, itp.
- Projektowanie wstępujące - od szczegółu do ogółu
- Projektowanie zstępujące

# Źródła danych

- Dane pochodzące z heterogenicznych źródeł – niezbędne procesy:
  - Integracja
  - Czyszczenie
  - Odświeżanie
- Problemy:
  - Niejednorodność danych
  - Niespójność danych
  - Brak danych
  - ...
- Zasilanie hurtowni - proces ETL – extract, transform, load

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Cele tworzenia hurtowni danych

- Wykonywanie analiz biznesowych bez ingerencji w systemy transakcyjne
- Wspomaganie decyzji (Decision Support Systems, Knowledge Discovery in Databases)
- Całościowy wgląd w dane firmy
- Dostęp do danych historycznych
- Ujednolicenie posiadanych informacji

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

## Typowe zastosowania

- Analiza trendów i zachowań
- Wykrywanie oszustw
- Ukierunkowany marketing
- Analiza rentowności
- Zapobieganie odejściu klienta
- Zarządzanie zasobami
- Automatyczne generowanie zamówień
- Analiza ryzyka kredytowego
- Długoterminowa ocena wartości klienta

# Funkcje i użytkownicy systemów analitycznych

## Funkcje systemów analitycznych:

- Zrozumieć (przedsięwzięcie, miary PKI)
- Wspomóc przy podejmowaniu decyzji
- Dostarczyć prognozy
- Wykryć trendy

## Użytkownicy systemów analitycznych:

- Dyrekcja i zarząd
- Analitycy, kontrolerzy, planiści
- Pracownicy wiedzy
- Aplikacje

# Założenia projektowe

## **Wymagania:**

- Jakie decyzje mają być wspomagane przez HD?
- Kto będzie podejmował decyzje?
- Czy istnieją różnice w znaczeniu miar?
- Jaki jest wymagany poziom interakcji użytkowników?
- Jakie opóźnienia są akceptowalne?

## **Ograniczenia:**

- Użyteczność dla użytkowników
- Koszty wdrożenia i zarządzania
- Zespół projektowy i administracyjny
- Techniczne protokoły i standardy
- Ciągły rozwój



*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Całkowity koszt posiadania hurtowni

- Planowanie i projektowanie
- Tworzenie i implementacja projektu
- Wdrożenie i zarządzanie
- Zakup i aktualizacja sprzętu
- Zakup lub tworzenie oprogramowania

**„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”**

# Pytania

- W jakich warunkach OLTP nie wystarcza?
- Czy w hurtowni gromadzimy i przetwarzamy wszystkie dostępne dane źródłowe? Dlaczego?
- Czy w systemach OLAP ważna jest normalizacja? Dlaczego?
- Z czego wynika trudność projektowania hurtowni danych?

## Quiz: HD czy OLTP?

- Ścisły podział na obszary informacyjne (tematyczne); lokalizacja danych zależy od tematyki
- Aplikacje projektowane są wokół procesów oraz funkcji, które wymagają różnych danych
- Jedno źródło, bez integracji
- Spójność danych, kluczy, synonimy, homonimy, rekordy logiczne...
- Dane są znakowane czasowo, szerszy horyzont czasowy np. 5 lat, historyzacja zmian
- Dane są znakowane czasowo w wąskim horyzoncie np. 30-90 dni
- Transakcje (insert, delete, update)
- Dane nie są modyfikowane przez użytkowników, odczyt danych
- Ewidencja transakcji, wspieranie operacyjnego poziomu zarządzania (również podejmowania decyzji na tym poziomie)
- Wspomaganie procesów podejmowania decyzji (raczej na poziomach: taktycznym i strategicznym)



„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”

# Odpowiedź

Aspekt	HD	OLTP
Tematycznie zorientowana	Ścisły podział na obszary informacyjne (tematyczne); lokalizacja danych zależy od tematyki	Aplikacje projektowane są wokół procesów oraz funkcji, które wymagają różnych danych
Zintegrowana	Spójność danych, kluczy, synonimy, homonimy, rekordy logiczne...	Jedno źródło, bez integracji
Chronologiczna	Dane są znakowane czasowo, szerszy horyzont czasowy np. 5 lat, historyzacja zmian	Dane są znakowane czasowo w wąskim horyzoncie np. 30-90 dni
Trwała	Dane nie są modyfikowane przez użytkowników, odczyt danych	Transakcje (insert, delete, update)
Cel	Wspomaganie procesów podejmowania decyzji (raczej na poziomach: taktycznym i strategicznym)	Ewidencja transakcji, wspieranie operacyjnego poziomu zarządzania (również podejmowania decyzji na tym poziomie)



Fundusze  
Europejskie  
Wiedza Edukacja Rozwój



Politechnika Wrocławska

Unia Europejska  
Europejski Fundusz Społeczny



*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Hurtownie danych

Dziękuję za uwagę

dr inż. Marcin Maleszka



Fundusze  
Europejskie  
Wiedza Edukacja Rozwój



Politechnika Wrocławskiego

Unia Europejska  
Europejski Fundusz Społeczny



*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Hurtownie danych

**Wielowymiarowy model danych - warstwa logiczna**

**dr inż. Marcin Maleszka**

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

## Hurtownia danych - definicja

**Hurtownia danych to:**

- **tematycznie zorientowana**
- **zintegrowana**
- **chronologiczna**
- **trwała**

kolekcja danych do wspomagania procesów podejmowania decyzji



„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”

## HD vs OLTP

Aspekt	HD	OLTP
Tematycznie zorientowana	Ścisły podział na obszary informacyjne (tematyczne); lokalizacja danych zależy od tematyki	Aplikacje projektowane są wokół procesów oraz funkcji, które wymagają różnych danych
Zintegrowana	Spójność danych, kluczy, synonimy, homonimy, rekordy logiczne...	Jedno źródło, bez integracji
Chronologiczna	Dane są znakowane czasowo, szerszy horyzont czasowy np. 5 lat, historyzacja zmian	Dane są znakowane czasowo w wąskim horyzoncie np. 30-90 dni
Trwała	Dane nie są modyfikowane przez użytkowników, odczyt danych	Transakcje (insert, delete, update)
Cel	Wspomaganie procesów podejmowania decyzji (raczej na poziomach: taktycznym i strategicznym)	Ewidencja transakcji, wspieranie operacyjnego poziomu zarządzania (również podejmowania decyzji na tym poziomie)

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

## HD - architektura

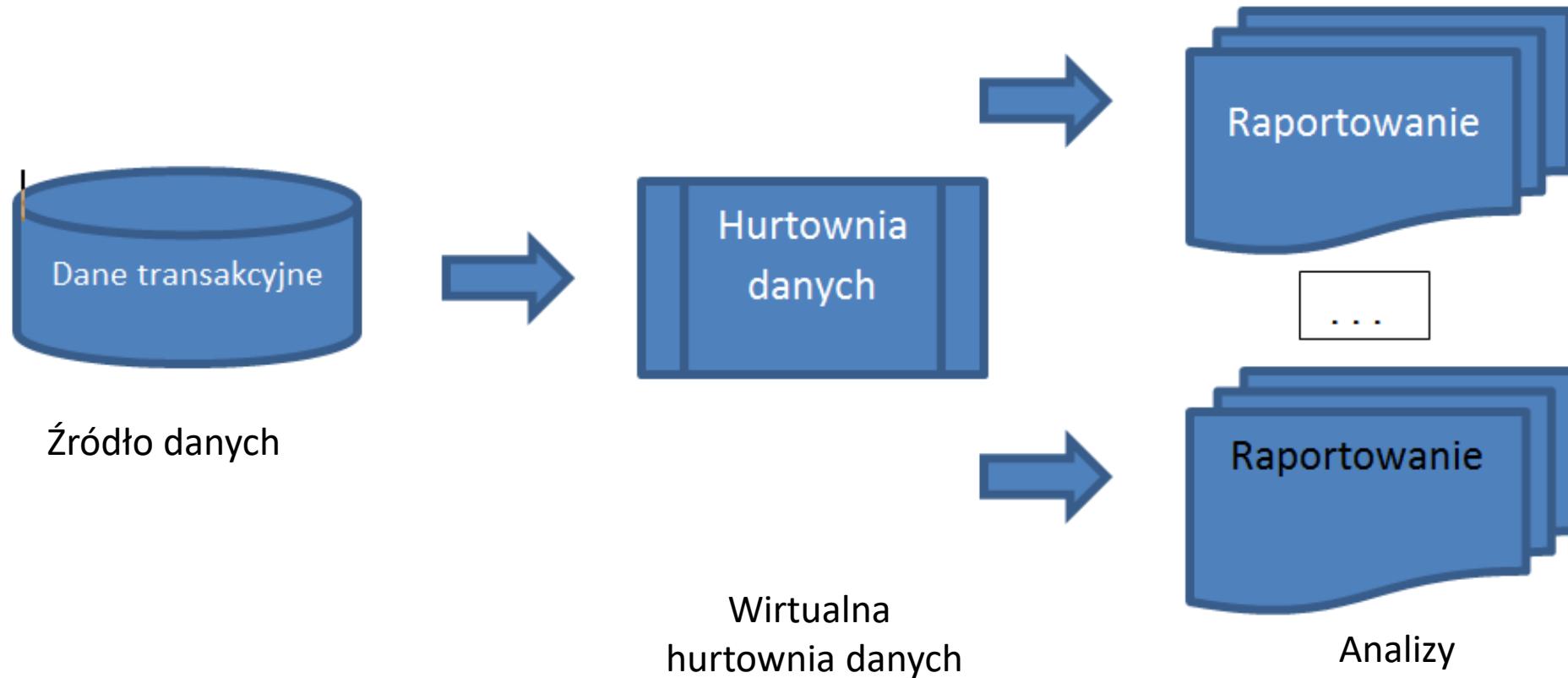
- Kluczowe właściwości architektury hurtowni danych:
  - Separacja – analityczne i transakcyjne procesy powinny być niezależnie realizowane bazując na rozłącznych zbiorach danych
  - Skalowalność – architektura sprzętowa i programowa powinna być łatwo modyfikowalna w kontekście zmieniających się wymagań użytkowników
- Podstawa klasyfikacji:
  - liczba warstw
  - liczba ról

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Architektura jednowarstwowa

- stosowana sporadycznie
- raporty ad hoc
- Zalety:
  - brak dodatkowych kosztów budowy i utrzymania HD
- Wady
  - trudności w tworzeniu zaawansowanych raportów
  - ograniczone możliwości analityczne

# Architektura jednowarstwowa





*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# **Wirtualna hurtownia danych**

- Implementowana jako widok danych transakcyjnych
- Brak separacji pomiędzy procesami transakcyjnymi i analitycznymi
- Zapytania tworzące informacje analityczne mają wpływ na regularnie realizowane operacje transakcyjne

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Modelowanie wielowymiarowe

- Analiza dziedziny problemowej
  - Identyfikacją i zrozumieniem procesów biznesowych
- Identyfikacja problemów i potrzeb w ramach rozpatrywanej dziedziny (biznesu)
- Ocena dostępności i jakości źródeł danych
- Określenie wymagań w kontekście ustalonych procesów i celów biznesowych

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

## Kluczowe etapy

1. Selekcja procesów biznesowych
2. Ustalenie poziomu szczegółowości (ziarnistości) rejestracji faktów
3. Identyfikacja faktów (zdarzeń biznesowych oraz wielkości pomiarowych istotnych w kontekście zarządzania i podejmowania decyzji)
4. Identyfikacja kontekstu (wymiarów) analizy faktów procesów biznesowych



*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Modelowanie konceptualne

1. Wyznaczenie procesów biznesowych
2. Określenie celu i zakresu analiz biznesowych w wybranych obszarach
3. Ocena dostępności i jakości źródeł danych
4. Zdefiniowanie modeli konceptualnych w kontekście uzgodnionych, wymaganych analiz faktów

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Modelowanie konceptualne

- Czy i jakie zbiory danych źródłowych są dostępne
  - gdzie wiadomo gdzie się znajdują dane
  - czy mamy dostęp do danych źródłowych?
- Kto w organizacji potrzebuje informacji udostępnianych w formie raportów?
- W jaki sposób można poprawić proces decyzyjny w zakresie krótko i dugo terminowym?
  - Więcej informacji
  - Udostępnić informację większej liczbie osób
  - Zmienić sposób dostępu do informacji

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Modelowanie konceptualne

- Jaki rodzaj informacji uznawany jest za potrzebny w procesie decyzyjnym?
- Czy są grupy osób, które nie mają dostępu do informacji lub dostęp jest ograniczony, a ma to wpływ na podejmowane decyzje?
- Czy mamy możliwość uzyskać odpowiedź na pytania w rodzaju:
  - Co jeśli?
  - Dlaczego tak/nie?
  - Czy można?



*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# **Wymiar – kontekst analizy biznesowej**

- Wymiar jest kontekstem analizy umożliwiając uzyskanie odpowiedzi na następujące kwestie:
  - Kto, co, gdzie, kiedy, dlaczego, jak?
  - Wymiary implementowane są z wykorzystaniem źródeł danych, które zawierają opisowe atrybuty specyficzne dla dziedziny problemowej (biznesu)
  - Atrybuty umożliwiają grupowanie i filtrowanie faktów

„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”

## Fakt – podstawa oceny

- Fakt jest zdarzeniem, który podlega pomiarowi i jest charakteryzowany za pomocą zbioru miar i ich wartości
- Miary ilościowo charakteryzują zdarzenie w biznesie
- Rekord reprezentujący **fakt** w hurtowni danych **pozostaje w relacji z fizycznie zarejestrowanym zdarzeniem**
- Spójność z zadeklarowaną ziarnistością



Fundusze  
Europejskie  
Wiedza Edukacja Rozwój



Politechnika Wrocławskiego

Unia Europejska  
Europejski Fundusz Społeczny



*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Miary

- wartości, które chcemy analizować (wartość sprzedaży, liczba pracowników, zadłużenie, zysk)
- wartości faktów świata rzeczywistego
- liczby (cecha addytywna lub semi-addytywna)
- podlegają ocenie wielowymiarowej

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Wymiar

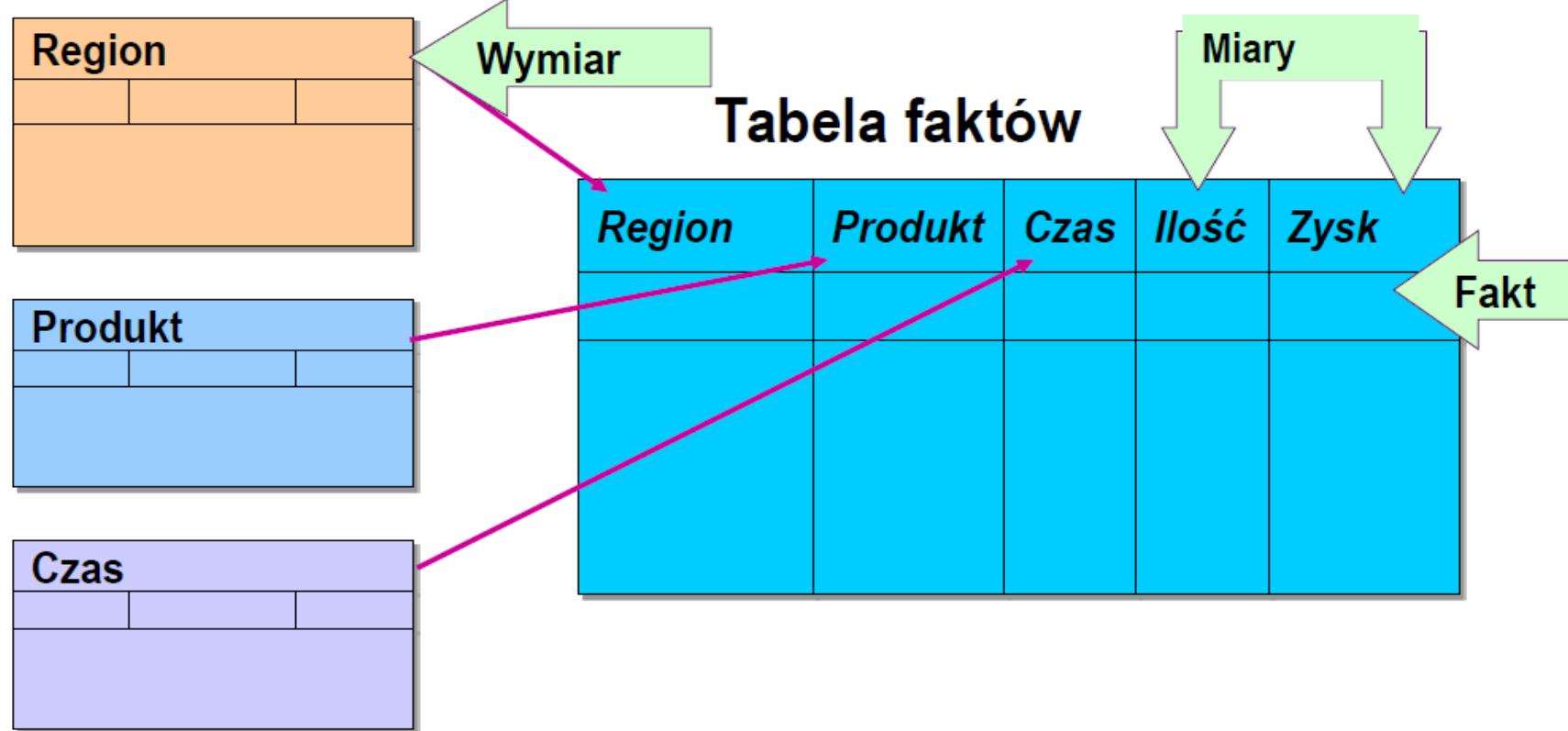
- Zbiór cech (atrybutów) istotnych z punktu widzenia analizy wartości faktów
- Wyznacza kontekst analizy wartości miar (region, produkt)
- Pozwala analizować informacje na różnych poziomach szczegółowości
- Ma charakter tekstowy, opisowy (klienci, regiony, daty).

„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”

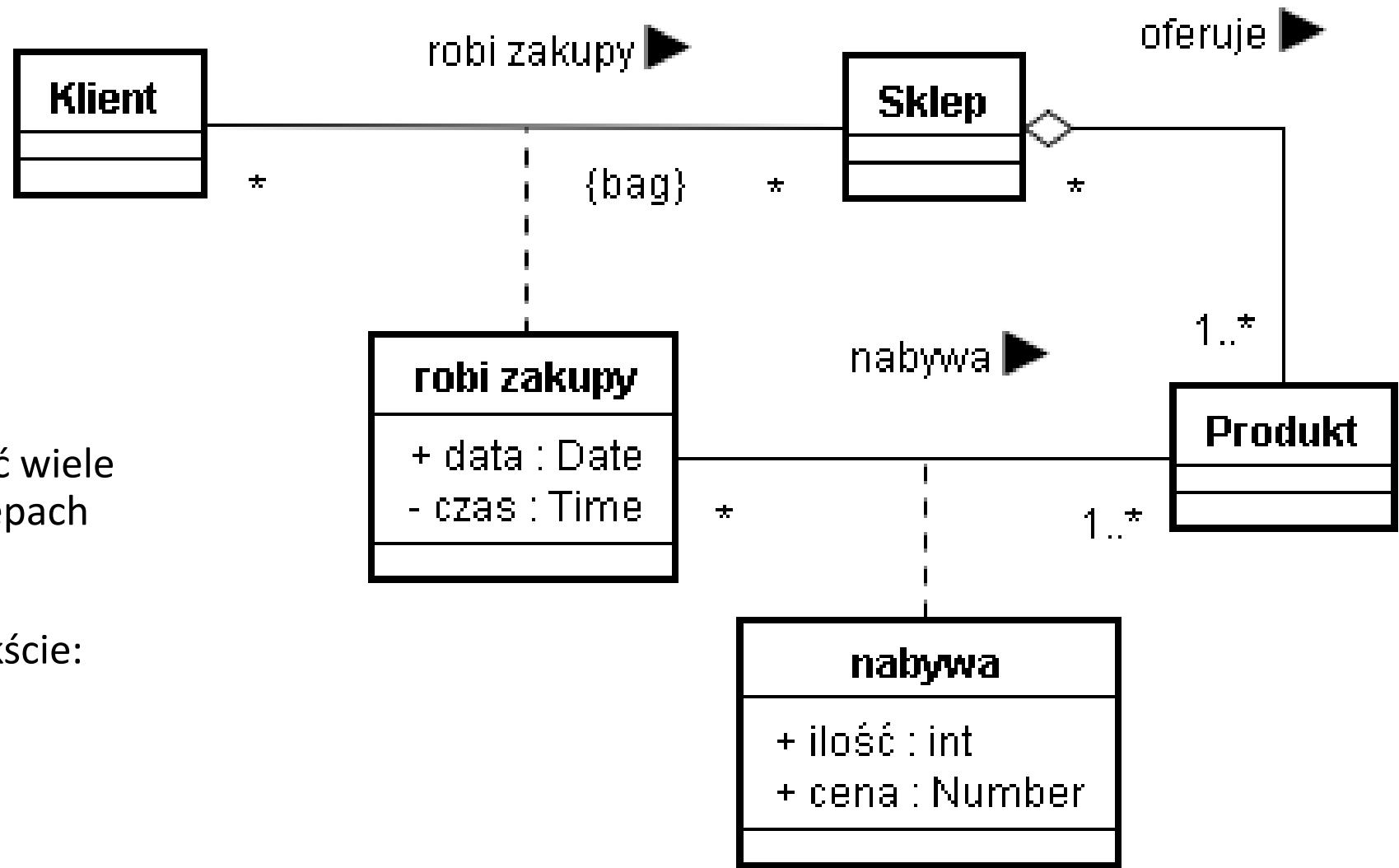
# Model wielowymiarowy

- Modele wielowymiarowe mogą być implementowane:
  - w bazach relacyjnych, lub
  - w bazach wielowymiarowych - analityczne modele OLAP (Online Analytical Processing Cube) - **kostki**
- Modele reprezentowane w bazie relacyjnej składają się z tabel **faktów** połączonych z tabelami **wymiarów** za pomocą kluczy obcych (tabele faktów) i kluczy głównych (tabele wymiarów)
- **Kostki** zawierają atrybuty wymiarów i faktów oraz wartości atrybutów na różnych poziomach agregacji

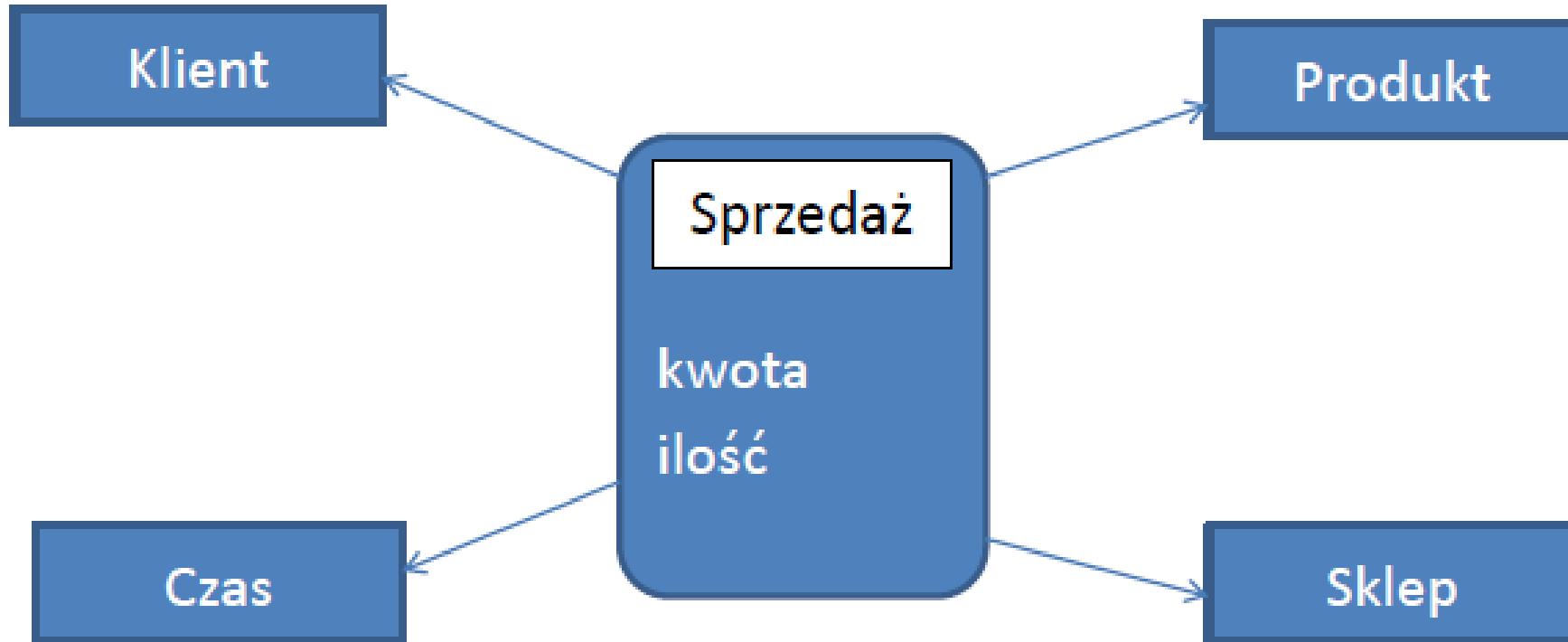
## Tabele wymiarów



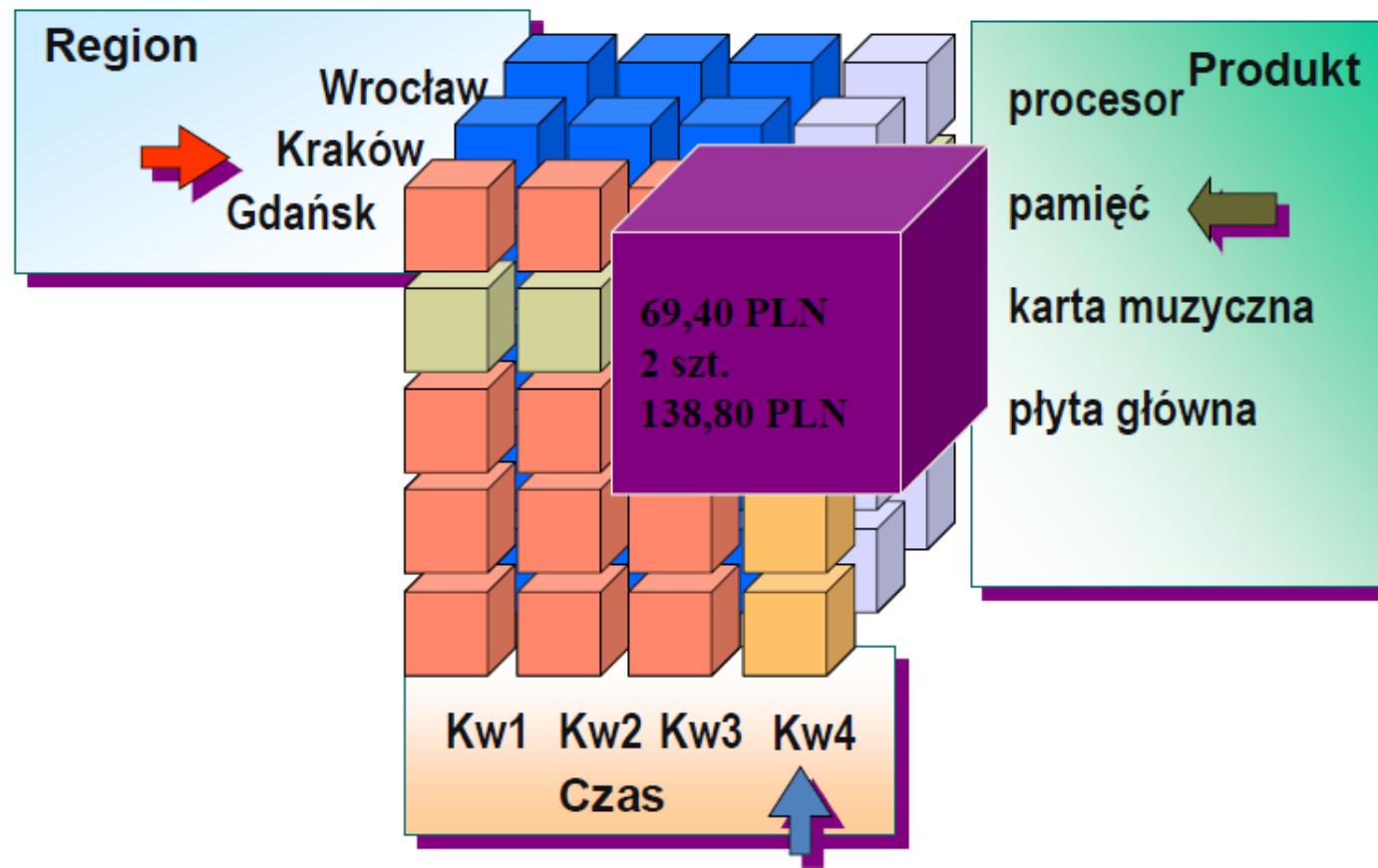
- wielu klientów może kupić wiele produktów w różnych sklepach
- ocena sprzedaży w kontekście:
  - klientów
  - sklepów
  - produktów
  - czasu



# Modelowanie konceptualne



## Przykład kostki





*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Wymiar zdegenerowany

- Jeżeli tabela faktów zawiera, oprócz kluczy obcych i miar, dodatkowe kolumny, to oznacza, że te kolumny pełnią funkcję zdegenerowanego wymiaru.
- Mogą to być naturalne klucze obiektów stanowiących kontekst analizy faktów.
- W przypadku wymiaru zdegenerowanego nie wykorzystuje się tabeli wymiarów.

## **Wymiar czasowy**

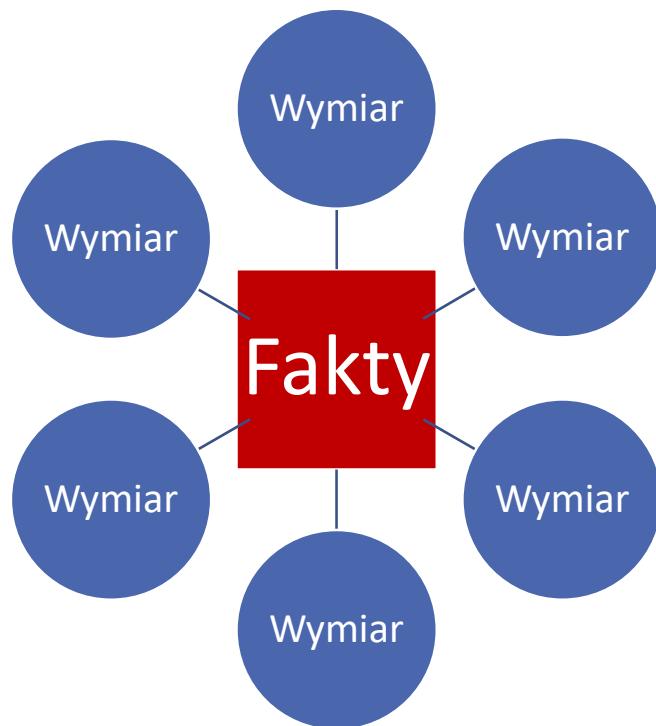
- umożliwia analizę biznesową w kontekście historii zdarzeń (faktów)
- często ma strukturę hierarchiczną:
  - rok – kwartał – miesiąc – dzień
- rejestracja czasu:
  - czas wykonania transakcji
  - dane historyczne
  - logi DBMS
  - porównywanie plików
  - ingerencja w system

# Wolno zmieniające się wymiary

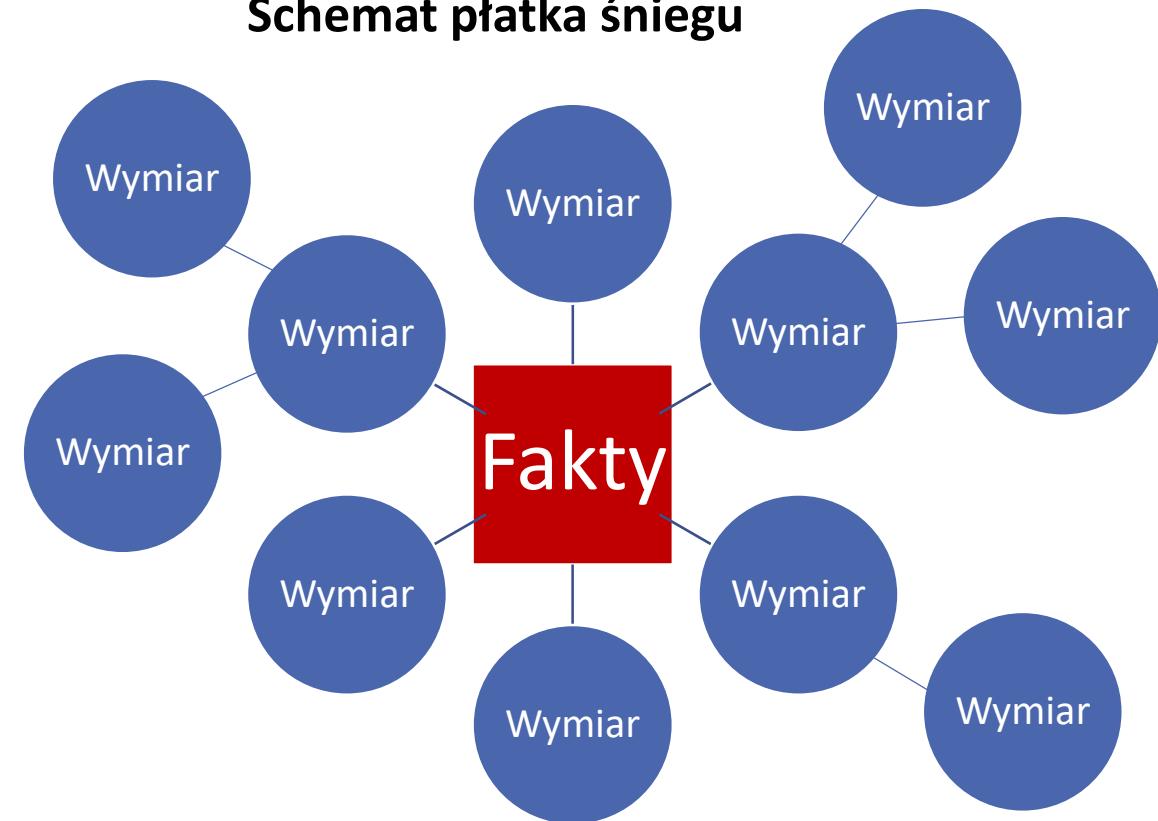
- Przyczyny:
  - Powiązanie pozycji wymiaru z faktem jest zmieniane lub anulowane
  - Wartości atrybutów pozycji wymiaru ulegają zmianie (w kontekście czasu) w rozpatrywanym wycinku rzeczywistości
- Typy:
  - Zmiana traktowana jest jako błąd (Typ 0)
  - Pamiętana jest ostatnia wartość (nadpisanie -Typ 1)
  - Pamiętana jest cała historia zmian (Typ 2)
  - Pozostawia się historię zmian w ograniczonym zakresie np. trzy ostatnie zmiany (Typ 3)

# Schematy modeli analitycznych

Schemat gwiazdy



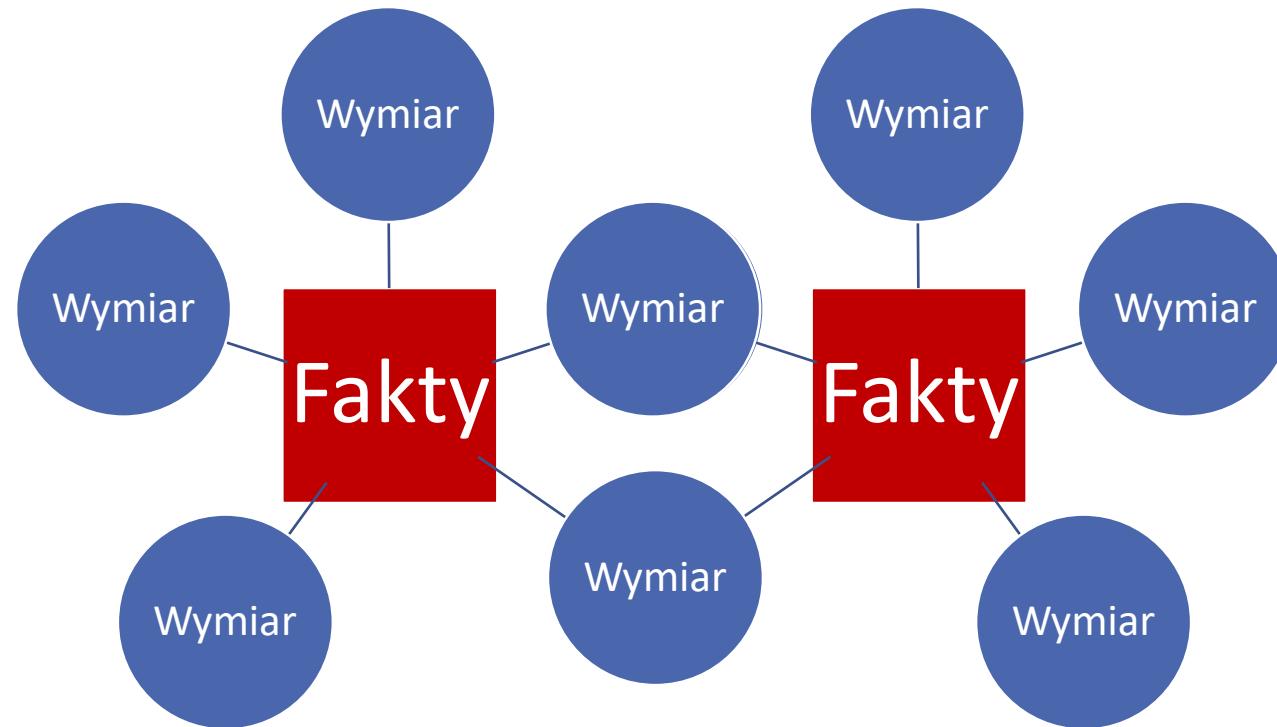
Schemat płatka śniegu



„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”

# Schematy modeli analitycznych

## Konstelacja faktów





Fundusze  
Europejskie  
Wiedza Edukacja Rozwój



Politechnika Wrocławska

Unia Europejska  
Europejski Fundusz Społeczny



*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Hurtownie danych

Dziękuję za uwagę

dr inż. Marcin Maleszka



Fundusze  
Europejskie  
Wiedza Edukacja Rozwój



Politechnika Wrocławska

Unia Europejska  
Europejski Fundusz Społeczny



*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Hurtownie danych

**Podstawy hurtowni danych**

**dr inż. Marcin Maleszka**

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Przypomnienie

- Czym jest hurtownia danych?
- Różnice pomiędzy systemami transakcyjnymi a analitycznymi
- Podstawowe pojęcia:
  - Fakt
  - Wymiar
  - Miara
  - Kostka

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# **Etapy projektowania hurtowni danych**

- Zrozumienie „potrzeby biznesu”
- Zrozumienie dziedziny problemowej
- Problemy w określonym wycinku rzeczywistości
- Identyfikacja potrzeb, celu i możliwości analiz biznesowych
- Wspieranie procesów decyzyjnych



*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

## Przykład – ogólna charakterystyka obszaru analizy

- Fabryka samochodów i podwykonawcy części aut.
- Produkcja dwóch marek samochodów – kilka modeli z każdej marki.
- Auta można nabyć jedynie za pośrednictwem dealerów.
- Dealerzy są rozliczani ze swojej sprzedaży miesięcznie/kwartalnie/itp.
- Funkcjonują wspólne oferty promocyjne dla całego obszaru sprzedaży.
- Każda fabryka podzespołów i każdy dealer operuje własnym sprzętem i oprogramowaniem.
- ...

## Przykład – obszar analizy

- Jaki jest miesięczny trend sprzedaży pod względem liczby i kwot sprzedawanych w dolarach każdej marki, modelu, serii i koloru (MMSC) dla konkretnego dealera, według każdego obszaru sprzedaży, regionu sprzedaży i stanu?
- Jaki jest wzorzec miesięcznej ilości zapasów według MMSC dla każdego dystrybutora, według każdego obszaru, regionu sprzedaży i stanu?
- Jak zmienia się miesięczna liczba sprzedanych samochodów ze względu na MMSC o określonym typie emisji - według dealera, fabryki, obszaru i regionu sprzedaży - w porównaniu z tymi samymi przedziałami czasowymi w poprzednim roku / poprzednich latach?
- Jaki jest trend w rzeczywistej miesięcznej sprzedaży (w dolarach i liczbach) MMSC dla każdego dystrybutora, obszaru i regionu sprzedaży w porównaniu do ich celów? Użytkownicy wymagają tych informacji zarówno według sum miesięcznych, jak i narastająco z roku na rok (YTD).
- Jaka jest historia (dwuletnie porównania) miesięcznej liczby jednostek sprzedawanych przez MMSC i powiązanych kwot w dolarach przez detalistów w porównaniu do hurtowników?

## Przykład – obszar analizy

- Jaka jest miesięczna sprzedaż według MMSC w tym roku w porównaniu do tego samego czasu w ubiegłym roku dla każdego dystrybutora?
- Jaki jest miesięczny trend według MMSC dla poszczególnych rodzajów promocji, według dealera, obszaru i regionu sprzedaży?
- Jaki jest miesięczny trend w średnim czasie, jaki zajmuje dealerowi sprzedaż określonej MMSC (zwanej prędkością i równą liczbą dni od otrzymania przez dealera samochodu do daty sprzedaży) według obszaru i regionu sprzedaży?
- Jaka była średnia miesięczna cena sprzedaży MMSC dla każdego dealera, obszaru i regionu sprzedaży?
- Jaki jest trend sprzedaży gotówkowych i kredytowych dla każdego dealera i rodzajów promocji na przestrzeni miesięcy i lat (porównać odpowiadające okresy sprzedaży)?
- Porównać miesięczne ceny sprzedaży i ilości od ostatniego modelu do bieżącego modelu nadwozia dla każdego regionu sprzedaży? Modele nadwozia zmieniają się co cztery lata.

## Przykład – obszar analizy

- Jaki jest miesięczny trend sprzedaży pod względem **liczby i kwot** sprzedawanych w dolarach każdej **marki, modelu, serii i koloru** (MMSC) dla konkretnego **dealera**, według każdego **obszaru sprzedaży, regionu sprzedaży i stanu**?
- Jaki jest wzorzec miesięcznej **ilości zapasów** według **MMSC** dla każdego **dystrybutora**, według każdego **obszaru, regionu sprzedaży i stanu**?
- Jak zmienia się miesięczna **liczba sprzedanych samochodów** ze względu na **MMSC** o określonym **typie emisji** - według **dealera, fabryki, obszaru i regionu sprzedaży** - w porównaniu z tymi samymi **przedziałami czasowymi** w poprzednim roku / poprzednich latach?
- Jaki jest trend w rzeczywistej **miesięcznej sprzedaży** (w dolarach i liczbach) **MMSC** dla każdego **dystrybutora, obszaru i regionu sprzedaży** w porównaniu do ich celów? Użytkownicy wymagają tych informacji zarówno według **sum miesięcznych**, jak i **narastająco z roku na rok** (YTD).
- Jaka jest historia (dwuletnie porównania) **miesięcznej liczby jednostek sprzedawanych** przez **MMSC** i powiązanych kwot w dolarach przez **detaelistów** w porównaniu do **hurtowników**?

## Przykład – obszar analizy

- Jaka jest **miesięczna sprzedaż** według **MMSC** w tym roku w porównaniu do tego samego **czasu** w ubiegłym roku dla każdego **dystributora**?
- Jaki jest **miesięczny trend** według **MMSC** dla poszczególnych **rodzajów promocji, według dealera, obszaru i regionu sprzedaży**?
- Jaki jest **miesięczny trend w średnim czasie**, jaki zajmuje **dealerowi** sprzedaż określonej **MMSC** (zwanej prędkością i równą liczbą dni od otrzymania przez dealera samochodu do daty sprzedaży) według **obszaru i regionu sprzedaży**?
- Jaka była **średnia miesięczna cena** sprzedaży **MMSC** dla każdego **dealera, obszaru i regionu sprzedaży**?
- Jaki jest trend **sprzedaży gotówkowych i kredytowych** dla każdego **dealera i rodzajów promocji** na przestrzeni **miesiący i lat** (porównać odpowiadające okresy sprzedaży)?
- Porównać **miesięczne ceny sprzedaży i ilości** od ostatniego modelu do bieżącego **modelu nadwozia** dla każdego **regionu sprzedaży**? Modele nadwozia zmieniają się co cztery lata.

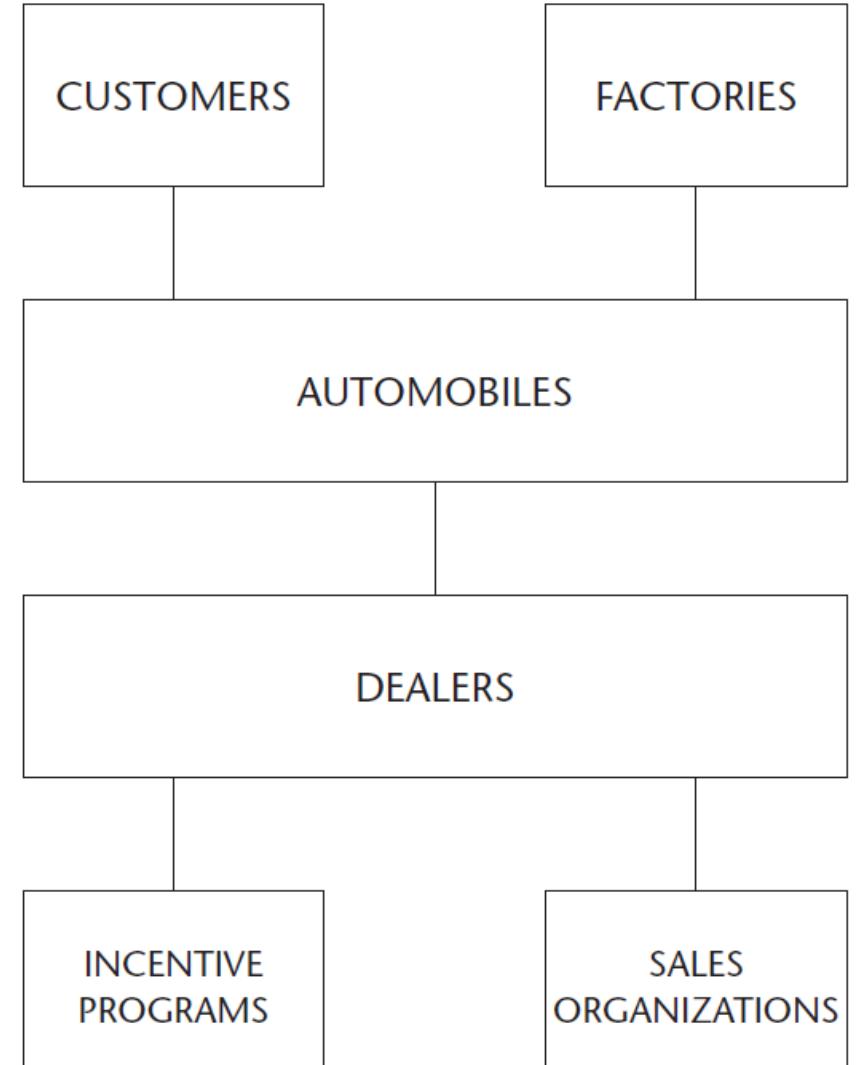
*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

## Zrozumienie „potrzeby biznesu”

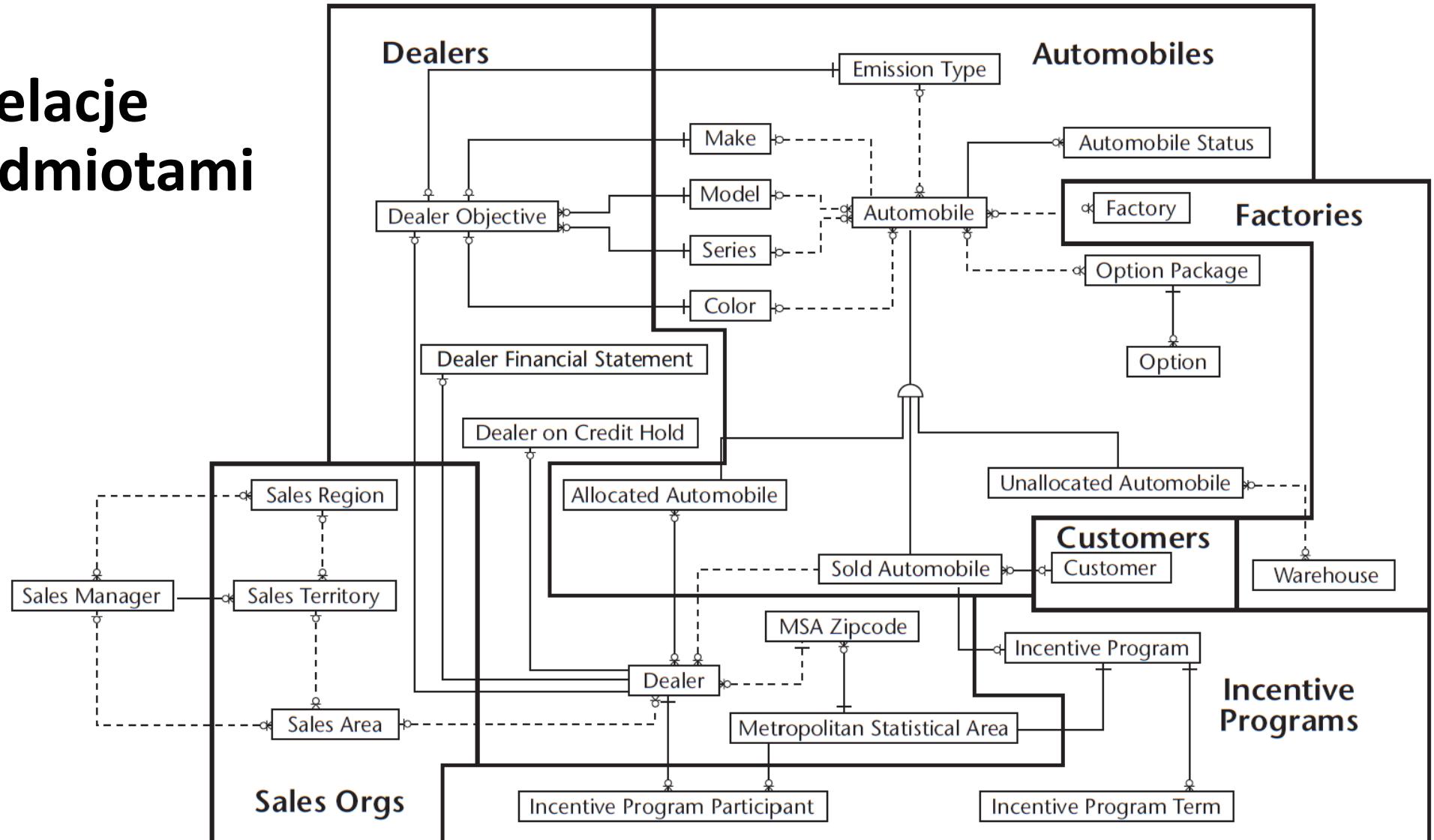
1. Zidentyfikuj obszar, z którego będą pobierane dane.
2. Zidentyfikuj interesujące podmioty w obszarze analizy i ustal ich identyfikatory.
3. Określ relacje pomiędzy tymi podmiotami.
4. Dodaj atrybuty.
5. Potwierdź strukturę modelu.
6. Potwierdź zawartość modelu.

# Przykład

1. Zidentyfikuj obszar – analiza sprzedaży aut
2. Zidentyfikuj interesujące podmioty w obszarze analizy i ustal ich identyfikatory:
  - Klient
  - Fabryka
  - Auto
  - Dealer
  - Promocja
  - Organizacja sprzedaży

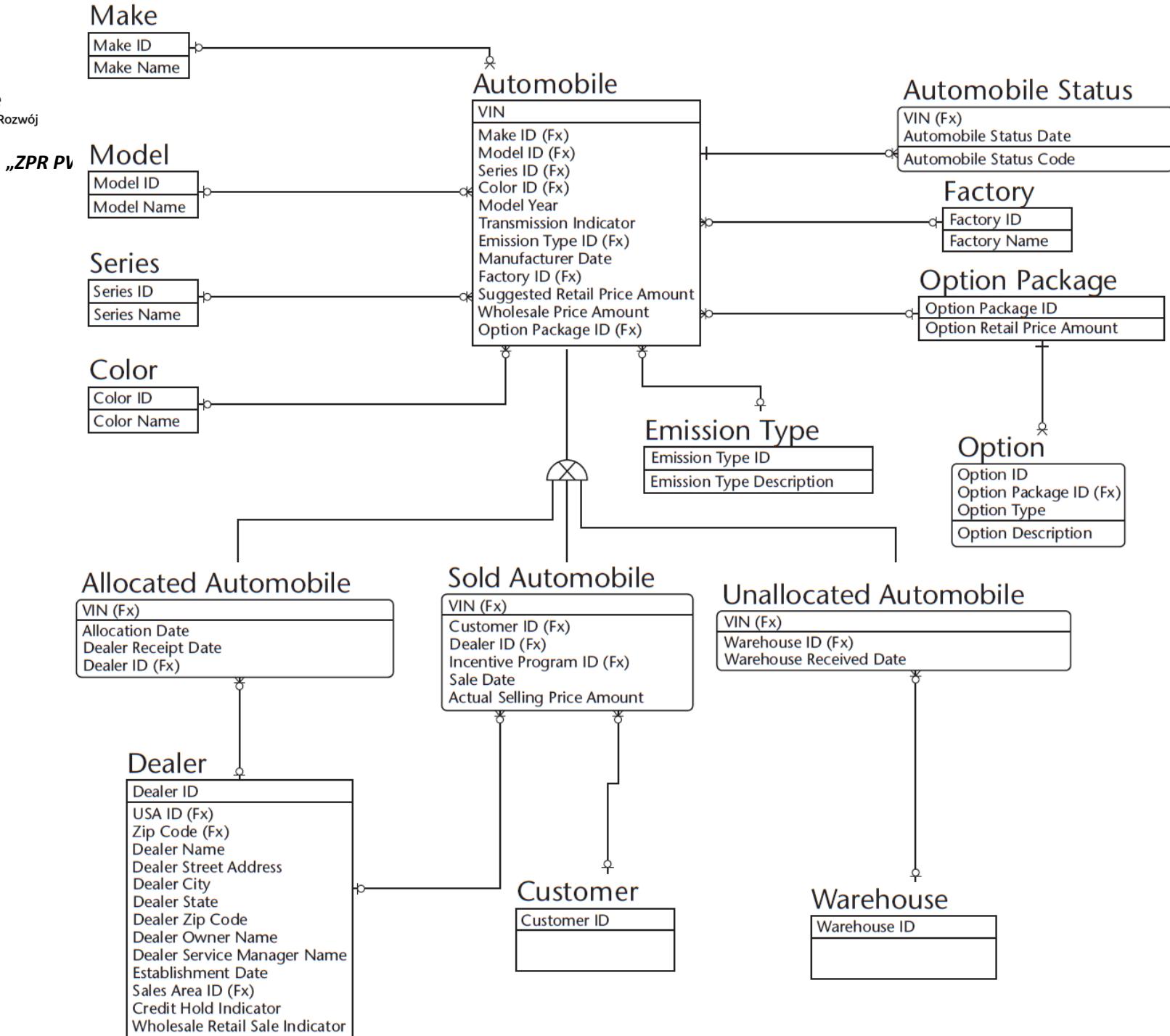


### 3. Określ relacje między podmiotami





## 4. Dodaj atrybuty



## Przykład cd.

### 5. Potwierdź strukturę modelu

- 3PN
- elastyczność, stabilność, spójność
- rzadko spotykana w wersji zaimplementowanej

### 6. Potwierdź zawartość modelu

- uzgodnienie poprawności z przedstawicielami biznesu
- sprawdzenie zgodności z regułami biznesowymi

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Zrozumienie dziedziny problemowej

- Biznesowy model danych -> model hurtowni danych
- Najważniejsze apsekty:
  - identyfikacja wymagań
  - wybór atrybutów
  - zapewnienie spójności danych
  - tworzenie widoków i targowisk danych
  - optymalizacja



Fundusze  
Europejskie  
Wiedza Edukacja Rozwój



Politechnika Wrocławskiego

Unia Europejska  
Europejski Fundusz Społeczny



„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”

# Metodologia tworzenia HD

## 1. Aspekty biznesowe:

1. Wybierz interesujące dane
2. Dodaj czas do klucza – perspektywa czasowa
3. Dodaj dane pochodne – zapewnienie spójności
4. Określ poziom ziarnistości

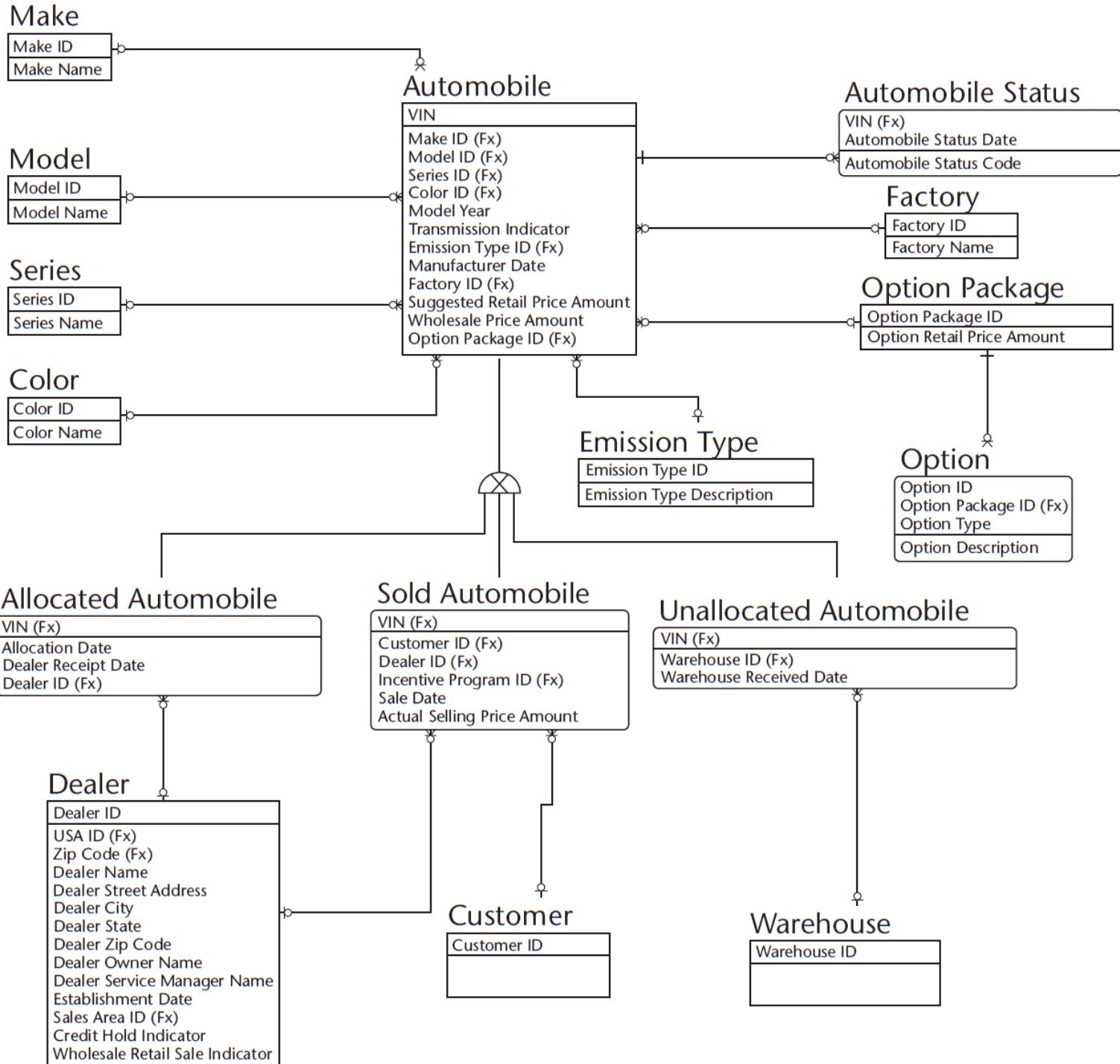
## 2. Aspekty wydajnościowe:

1. Dodaj podsumowania – w zależności od ustalonego poziomu szczegółowości
2. Dokonaj niezbędnych złączeń tabel
3. Utwórz tabele wymiarów i faktów
4. Segreguj dane – optymalizacja zapytań



# Dyskusja

- Zgodność modelu z wymaganiami biznesowymi?
- Problemy?
- Kwestie do wyjaśnienia?





„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”

# Podsumowanie - ćwiczenie

Krok	Cel	Opis
Wybór danych		
Czas w kluczu		
Dane powiązane		
Ziarnistość		
Podsumowania		
Złączenia tabel źródłowych		
Tworzenie tabel wynikowych		
Segregacja		

# **Etapy projektowania hurtowni danych**

1. Charakterystyka dziedziny problemowej
2. Krótki opis obszaru analizy
3. Problemy i potrzeby
4. Cel przedsięwzięcia
  1. Oczekiwania
  2. Zakres analizy
5. Źródła danych (lokalizacja, format, dostępność)
6. Wstępna analiza źródeł danych

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Wstępna analiza źródeł danych

1. Profilowanie danych
  1. Analiza danych
  2. Ocena przydatności danych w pliku do tworzenia hurtowni danych
2. Definicja typów encji/klas (wraz z własnościami) oraz związków pomiędzy nimi
3. Propozycja wymiarów, hierarchii, miar (w tym nieaddytywnych)
4. Model konceptualny
5. Implementacja bazy danych

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

## Tworzenie kluczy do tabel

- Potencjalne problemy:
  - niespójność – przykłady?
  - unikalność wartości
- Atrybuty będące potencjalnymi kandydatami na klucz:
  - istniejące w systemie
  - uznane standardowe klucze
  - klucze sztuczne

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

## **Wymiar czasowy**

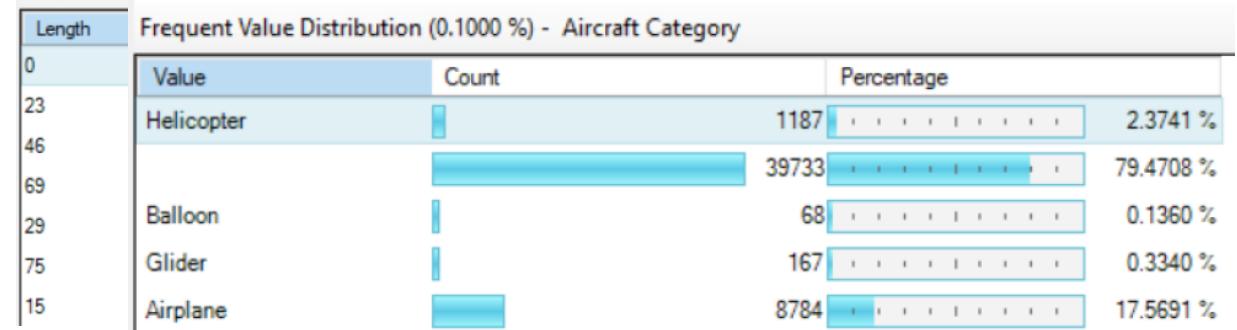
- Różne rodzaje kalendarza
- Różne długości miesięcy
- Obsługa dni wolnych
- Często nadmiarowo przechowywane atrybuty:
  - Nr dnia w roku (w roku fiskalnym)
  - Nr dnia w miesiącu
  - Nr miesiąca
  - Nazwa miesiąca (różne wersje językowe)
  - Nazwa dnia tygodnia (różne wersje językowe)
  - Data początku i końca tygodnia
  - Nr kwartału

# Profilowanie danych

Key Columns	Key Strength
Accident Number	..... 100.0000 %
Event Id	..... 98.5839 %

- Klucze kandydujące
  - Procent brakujących danych
  - Rozkład długości ciągów znakowych w danych
  - Rozkład częstości występowania wartości
  - Wartości unikalne

Column	Column	Number Of Distinct Values
Accident Nu		1
Air Carrier	Accident Number	49997
Aircraft Cat	Air Carrier	1744
Aircraft Dan	Aircraft Category	10
Airport Code	Aircraft Damage	4
Airport Nam	Airport Code	7462
Amateur Bui	Airport Name	14167
Broad Phas	Amateur Built	3
Country	Broad Phase of Flight	13
< [ ] >	Country	170





Fundusze  
Europejskie  
Wiedza Edukacja Rozwój



Politechnika Wrocławska

Unia Europejska  
Europejski Fundusz Społeczny



*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Hurtownie danych

Dziękuję za uwagę

dr inż. Marcin Maleszka



Fundusze  
Europejskie  
Wiedza Edukacja Rozwój



Politechnika Wrocławskiego

Unia Europejska  
Europejski Fundusz Społeczny



*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Hurtownie danych

**Podstawy procesu ETL**

**dr inż. Marcin Maleszka**

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

## Hurtownia danych - definicja

**Hurtownia danych to:**

- **tematycznie zorientowana**
- **zintegrowana**
- **chronologiczna**
- **trwała**

kolekcja danych do wspomagania procesów podejmowania decyzji

# ETL

- Extract
  - Transform
  - Load
- 
- Zastosowanie reguł biznesowych do istniejących danych w celu uzyskania użytecznych informacji
  - Czyszczenie i standaryzacja danych
  - Integracja różnych danych (wewnętrznych i zewnętrznych)
  - Agregacja danych
  - Nawet 70% - 80% wysiłku budowy hurtowni danych

# ETL

- Pobierz dane ze źródła i załaduj do hurtowni
  - kopiowanie danych pomiędzy bazami
- Dane są wyciągane z bazy OLTP, przekształcane tak, aby pasowały do schematu hurtowni i ładowane do hurtowni
- Źródłowe dane mogą nie być przechowywane w bazie
- Myśl o procesie ETL, a nie o fizycznej implementacji tego procesu!

# ETL

- Złożona kombinacja procesu i technologii wymagająca nakładów sił i energii:
  - analityków biznesowych
  - projektantów baz danych
  - developerów aplikacji
- Nie myić procesu ETL z jednorazowym czy nawet okresowym dodawaniem danych do bazy!
- Proces:
  - zautomatyzowany
  - udokumentowany
  - łatwo modyfikowalny

# Extraction

- integracja wszystkich systemów przedsiębiorstwa
- heterogeniczne źródła danych
- każde źródło danych ma swoją charakterystykę:
  - DBMS
  - system operacyjny
  - hardware
  - protokoły komunikacji
- Logiczna mapa danych
  - określa relacje pomiędzy skrajnymi etapami procesu ETL

# Extraction

Cel		Źródło			Przekształcenie	
Tabela	Kolumna	Typ danych	Tabela	Kolumna	Typ danych	

- dokładnie wiadomo, co dzieje się z danymi
- przekształcenie – zazwyczaj SQL



„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”

# Mapa logiczna

Target					Source				Transformation
Table Name	Column Name	Data Type	Table Type	SCD Type	Database Name	Table Name	Column Name	Data Type	
EMPLOYEE_DIM	EMPLOYEE_KEY	NUMBER	Dimension	1				NUMBER	Surrogate key.
EMPLOYEE_DIM	EMPLOYEE_ID	NUMBER	Dimension	1	HR_SYS	EMPLOYEES	EMPLOYEE_ID	NUMBER	Natural Key for employee in HR system
EMPLOYEE_DIM	BIRTH_COUNTRY_NAME	VARCHAR2(75)	Dimension	1	HR_SYS	COUNTRIES	NAME	VARCHAR2(75)	select cname from employees e, states s, countries c where e.state_id = s.state_id and s.country_id = c.country
EMPLOYEE_DIM	BIRTH_STATE	VARCHAR2(75)	Dimension	1	HR_SYS	STATES	DESCRIPTION	VARCHAR2(255)	select s.description from employees e, states s where e.state_id = s.state_id
EMPLOYEE_DIM	DISPLAY_NAME	VARCHAR2(75)	Dimension	1	HR_SYS	EMPLOYEES	FIRST_NAME	VARCHAR2(75)	select initcap(salutation)    ' '    initcap(first_name)    ' '    initcap(last_name) from employee
EMPLOYEE_DIM	BIRTH_DATE	DATE	Dimension	1	HR_SYS	EMPLOYEES	DOB	DATE	trunc(DOB)
EMPLOYEE_DIM	SALUTATION	VARCHAR2(12)	Dimension	1	HR_SYS	EMPLOYEES	SALUTATION	VARCHAR2(12)	initcap(salutation)
EMPLOYEE_DIM	FIRST_NAME	VARCHAR2(30)	Dimension	1	HR_SYS	EMPLOYEES	FIRST_NAME	VARCHAR2(30)	initcap(first_name)
EMPLOYEE_DIM	LAST_NAME	VARCHAR2(30)	Dimension	1	HR_SYS	EMPLOYEES	LAST_NAME	VARCHAR2(30)	initcap(last_name)
EMPLOYEE_DIM	MARITAL_STATUS	VARCHAR2(12)	Dimension	2	HR_SYS	MARITAL_STATUS	DESCRIPTION	VARCHAR2(12)	select initcap(m.name,'Unknown') from employee e, marital_status m where e.marital_status_id = m.marital_status_id
EMPLOYEE_DIM	DIVERSITY_CATEGORY	VARCHAR2(30)	Dimension	1	HR_SYS	EMPLOYEES	EEO_CLASS	VARCHAR2(30)	decode(eeo_class, null, 'Not Stated', decode(eeo_class, 'N', 'Not Stated', eeo_class))
EMPLOYEE_DIM	GENDER	VARCHAR2(12)	Dimension	1	HR_SYS	EMPLOYEES	SEX	VARCHAR2(12)	initval(sex, 'Unknown')
EMPLOYEE_DIM	EMPLOYEE_STATUS	VARCHAR2(24)	Dimension	1	HR_SYS	EMPLOYEES	STATUS	VARCHAR2(24)	select es.name from employee e, employee_statuses es where e.employee_status_id = es.employee_status_id
EMPLOYEE_DIM	POSITION_CODE	VARCHAR2(12)	Dimension	2	HR_SYS	POSITIONS	POSITION_CODE	VARCHAR2(12)	select p.code from employees e, positions p where e.position_id = p.position_id
EMPLOYEE_DIM	POSITION_CATEGORY	VARCHAR2(30)	Dimension	2	HR_SYS	POSITIONS	POSITION_CATEGORY	VARCHAR2(30)	select p.category from employees e, positions p where e.position_id = p.position_id
EMPLOYEE_DIM	HIRE_DATE	DATE	Dimension	1	HR_SYS	EMPLOYEES	DATE_HIRED	DATE	trunc(date_hired)

# Fazy ekstrakcji

## 1. Wykrywanie danych:

- czystość danych
- spójność danych
- identyfikacja i sprawdzenie źródła pod kątem założonego celu
  
- dokumentacja systemu źródłowego
- śledzenie zmian w systemie
- określenie miejsca pochodzenia danych
- świadomość redundancji danych (dane kopiowane, przekształcane, czyszczone, itp.)



*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Fazy ekstrakcji

## 2. Detekcja anomalii:

- NULL (operacjełączenia tabel)
- wartości kluczowe
- daty
- audit columns – używane przez DB, warunkowo uaktualniane

## 3. Eliminacja anomalii:

- tworzenie dwóch tabel (dane z poprzedniego i bieżącego ładowania)
- obliczanie różnicy pomiędzy tabelami w celu wykrycia zmian

# Transformation

- udokumentowany etap modyfikacji danych do pożądanej postaci
- paradygmaty jakości danych:
  - poprawność
  - jednoznaczność
  - spójność
  - kompletność
- dwukrotne sprawdzenie:
  - po ekstrakcji
  - po czyszczeniu i potwierdzeniu dodatkowych warunków

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Transformation - Czyszczenie danych

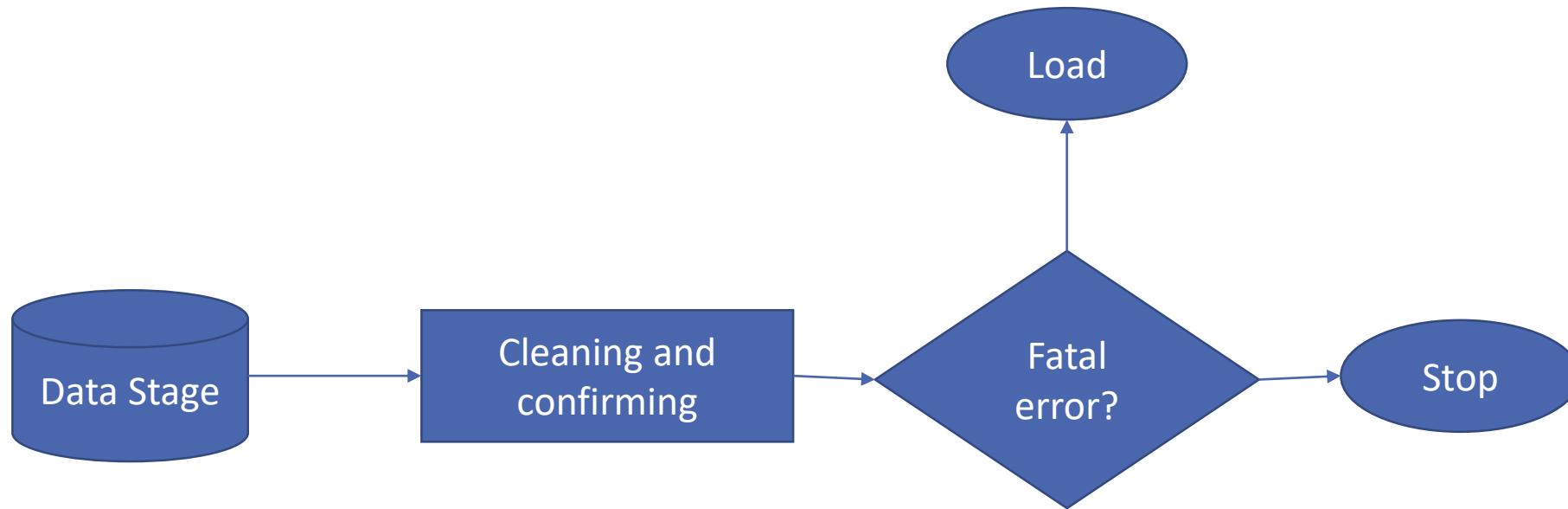
- detekcja anomalii:
  - próbkowanie danych
  - zliczanie rekordów
- sprawdzenie własności kolumn:
  - wartości NULL w miejscu kluczy
  - wartości numeryczne poza oczekiwany zakresem
  - zbyt długie/krótkie długości danych
  - dane poza zakresem zbioru
  - dane odstające od wzorca

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Transformation - potwierdzenie

- Sprawdzenie struktury
  - klucze główne i obce
  - integralność referencyjna kluczy
- Sprawdzenie danych i reguł
  - prostych reguł biznesowych
  - na poziomie logicznym

# Transformation



*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Loading

## Ładowanie danych do wymiarów

- minimalizacja zbioru komponentów
- prosty klucz główny
- denormalizacja tabel
  
- slowly changing dimensions
  - zapis wymiaru jako fizycznej tabeli na dysku
  
- przypisanie kluczy zastępczych



*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Loading

## Ładowanie danych do tabeli faktów

- w tabeli faktów przechowywane są miary
- uproszczone relacje pomiędzy tabelą faktów a wymiarami
- tworzenie klucza tabeli faktów
  - tworzenie klucza zastępczego

# ETL – zasilanie hurtowni danych danymi

- **ekstrakcja danych** z systemów źródłowych (SAP, ERP, inne systemy transakcyjne), dane z różnych systemów są konwertowane do wspólnego, jednolitego formatu hurtowni danych
- **transformacja danych:**
  - zastosowanie logiki biznesowej,
  - czyszczenie danych,
  - filtrowanie,
  - rozdzielenie jednej kolumny na kilka i odwrotnie,
  - łączenie danych z kilku źródeł (lookup, merge),
  - transpozycje kolumn i wierszy,
  - odrzucanie danych niespełniających pewnych zdefiniowanych
- **załadowanie danych** do hurtowni danych lub repozytoriów danych innych aplikacji raportujących

# Narzędzia do ETL

- Informatica - Power Center
- IBM - Websphere DataStage
- SAP - BusinessObjects Data Integrator
- IBM - Cognos Data Manager
- **Microsoft - SQL Server Integration Services**
- Oracle - Data Integrator (przed zakupem produkt Data Conductor firmy Sunopsis)
- SAS - Data Integration Studio
- SAS Viyo
- Oracle - Warehouse Builder
- AB Initio
- Information Builders - Data Migrator
- Pentaho - Pentaho Data Integration
- Embarcadero Technologies - DT/Studio
- IKAN - ETL4ALL
- IBM - DB2 Warehouse Edition
- Pervasive - Data Integrator
- ETL Solutions Ltd. - Transformation Manager
- Group 1 Software (Sagent) - DataFlow
- Sybase - Data Integrated Suite ETL
- Talend - Talend Open Studio
- Expressor Software - Expressor Semantic Data Integration System
- Elixir - Elixir Repertoire
- OpenSys - CloverETL

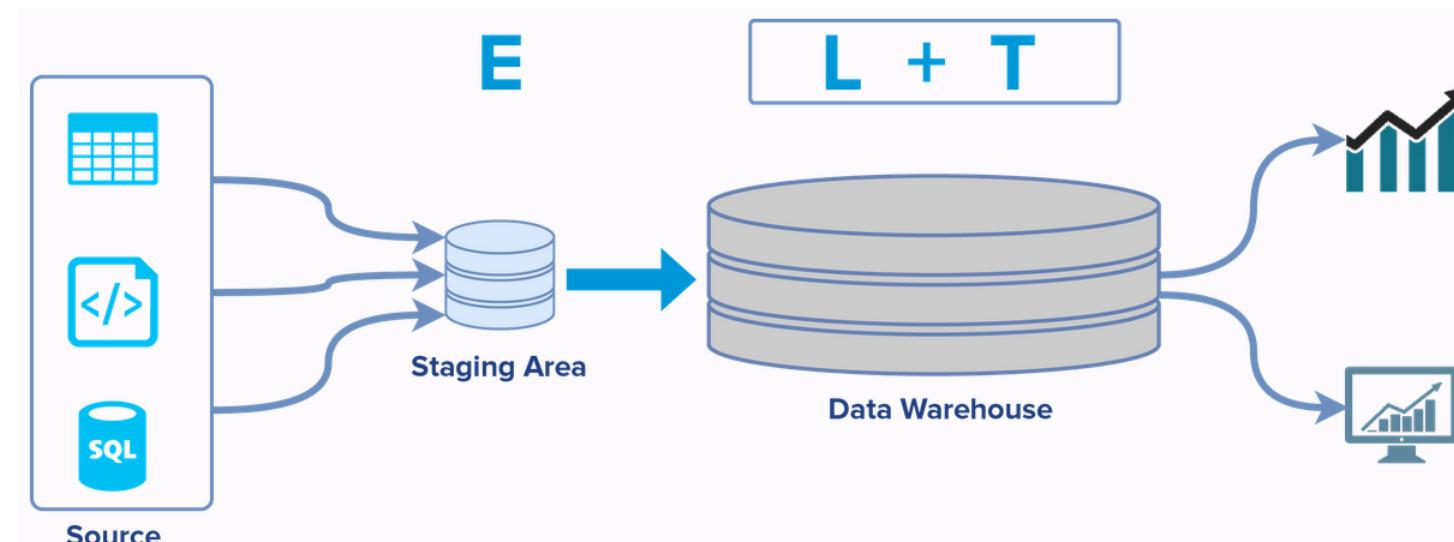
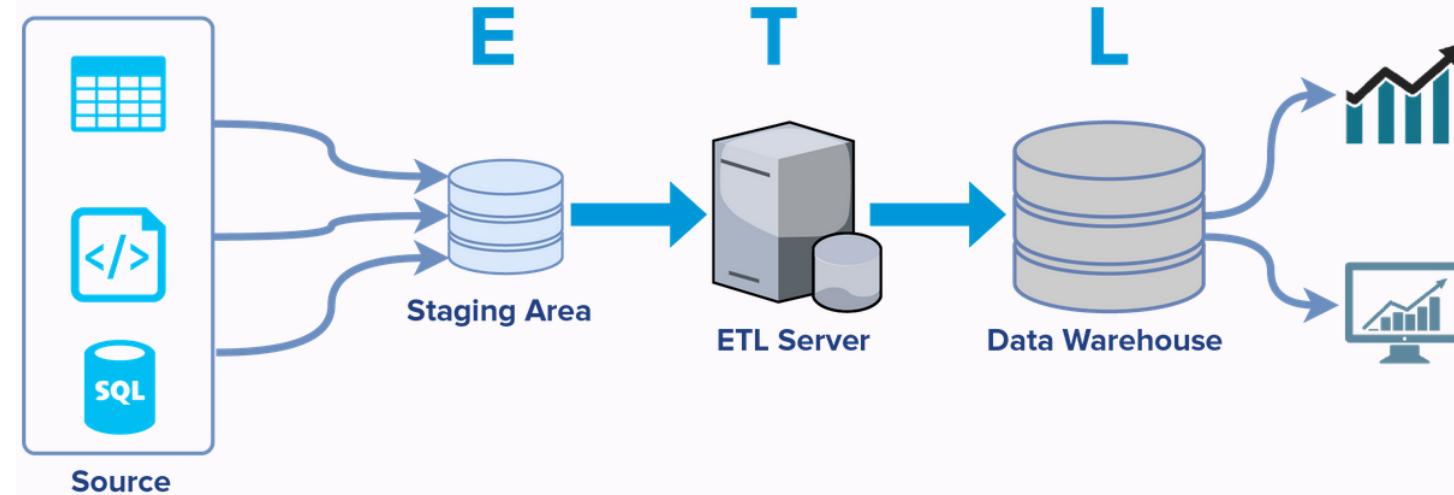
# ELT

- Dane ekstraktowane z systemów źródłowych bezpośrednio ładowane w oryginalnym formacie do bazy danych hurtowni danych
- Przy pomocy wygenerowanych poleceń i procedur SQL serwer bazy danych (DBMS) wykonuje transformacje danych
- Zasila tabele docelowe hurtownie
- Wymagania:
  - bardzo wydajny
  - wysoce skalowalny
  - i dobrze dostrojony serwer DBMS
- Stosowany przy bardzo dużych wolumenach danych



„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”

## ETL vs ELT



# ETL vs ELT

## ETL

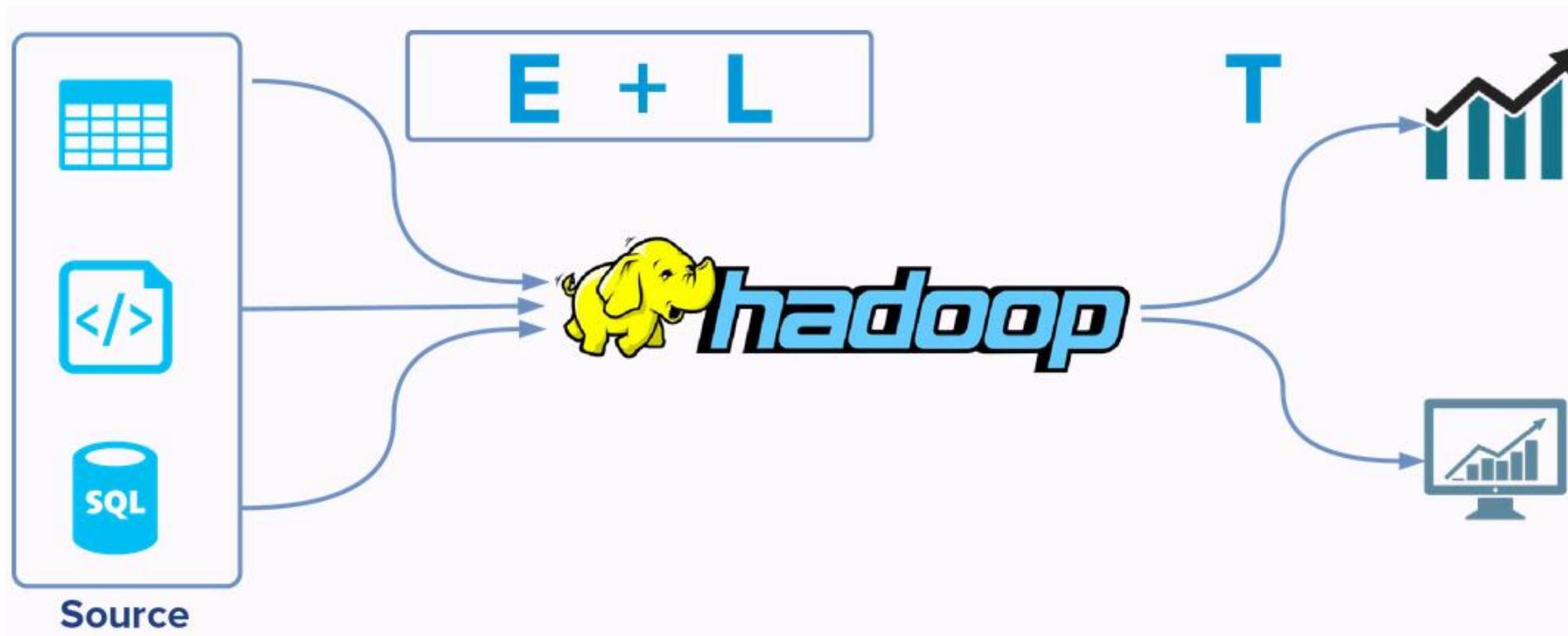
- Extract – wyładowanie danych i załadowanie ich do przestrzeni tymczasowej (ang. staging)
- Wada: niezbędny serwer na potrzeby narzędzia SQL
- Transform – przygotowanie modelu i przekształcenie danych do pożądanej postaci (ang. schema-on-write)
- Load

## ELT

- Extract – przygotowanie danych, ale bez definiowania, jak mają wyglądać dane wyjściowe (ang. schema-on-read)
- Load – załadowanie surowych danych do centralnego repozytorium danych (ang. Data Lake)
- Transform - wykorzystanie technologii pozwalającej przetwarzać dane nierelacyjne, w różnych formatach i strukturach

„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”

## Przykład ELT – Big Data



## Zalety i wady

Kryterium	ETL	ELT
Schemat	Podczas tworzenia hurtowni.	ELT nie wyklucza podejścia Schema-on-Write. Decyzja o formie danych podczas ich odczytu z repozytorium danych.
Zmiany w modelu hurtowni	Często musimy zmienić przepływ ETL oraz model hurtowni.	Zmiana może ograniczyć się do warstwy hurtowni danych i kroku transformacji.
Infrastruktura	Potrzebne dodatkowe maszyny.	Całość procesu realizowana na docelowym wystarczająco wydajnym serwerze.
Kompetencje	Wymagane dodatkowe kompetencje związane z procesami i narzędziami ETL.	L+T -> znajomość baz danych. W pozostałych przypadkach wymagana jest znajomość technologii, wykorzystywana do przechowywania i procesowania danych.



„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”

## Zalety i wady

Kryterium	ETL	ELT
<b>Czas dostępu do danych</b>	Zazwyczaj dane dostępne po ukończeniu całego procesu.	Dane szybciej dostępne na docelowej maszynie. Możemy mieć dostęp do danych surowych przed transformacją.
<b>Zastosowanie</b>	Rozwiązanie popularne i optymalne przy dużych wolumenach danych oraz skomplikowanych transformacjach.  Może nie być optymalne kosztowo dla małych rozwiązań.	Zysk widoczny przy przetwarzaniu potężnych zbiorów danych opartych o rozwiązania nastawione na skalowalność oraz dane nieustrukturyzowane.



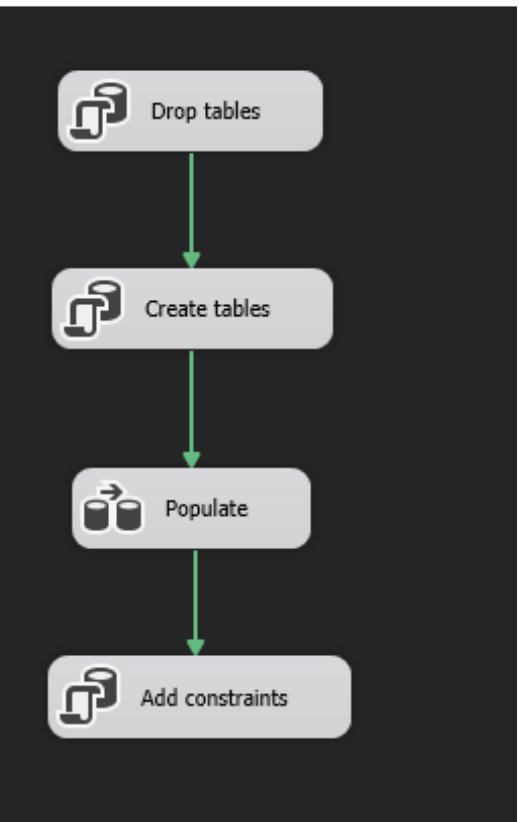
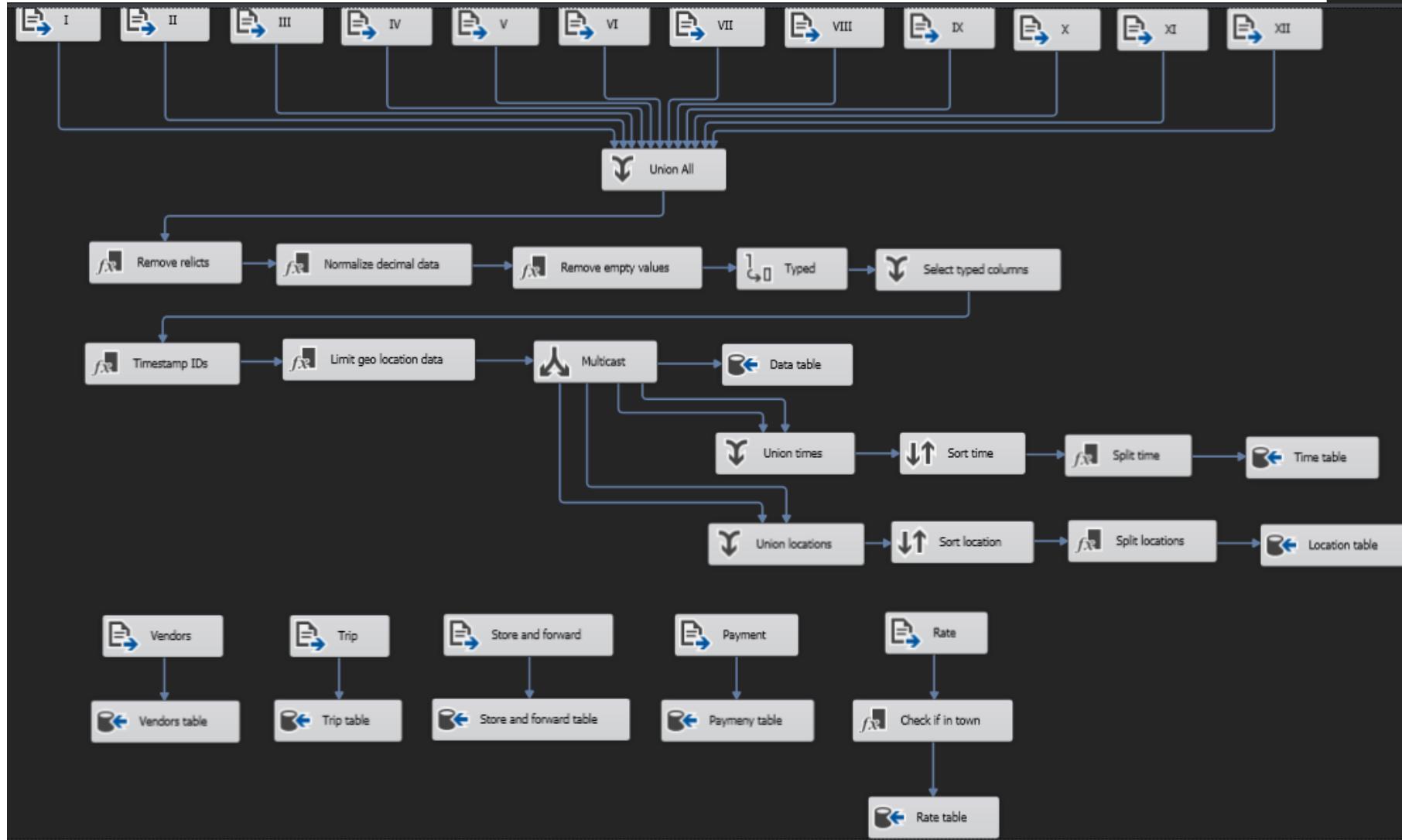
Fundusze  
Europejskie  
Wiedza Edukacja Rozwój



Politechnika Wrocławska

Unia Europejska  
Europejski Fundusz Społeczny

### „ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”



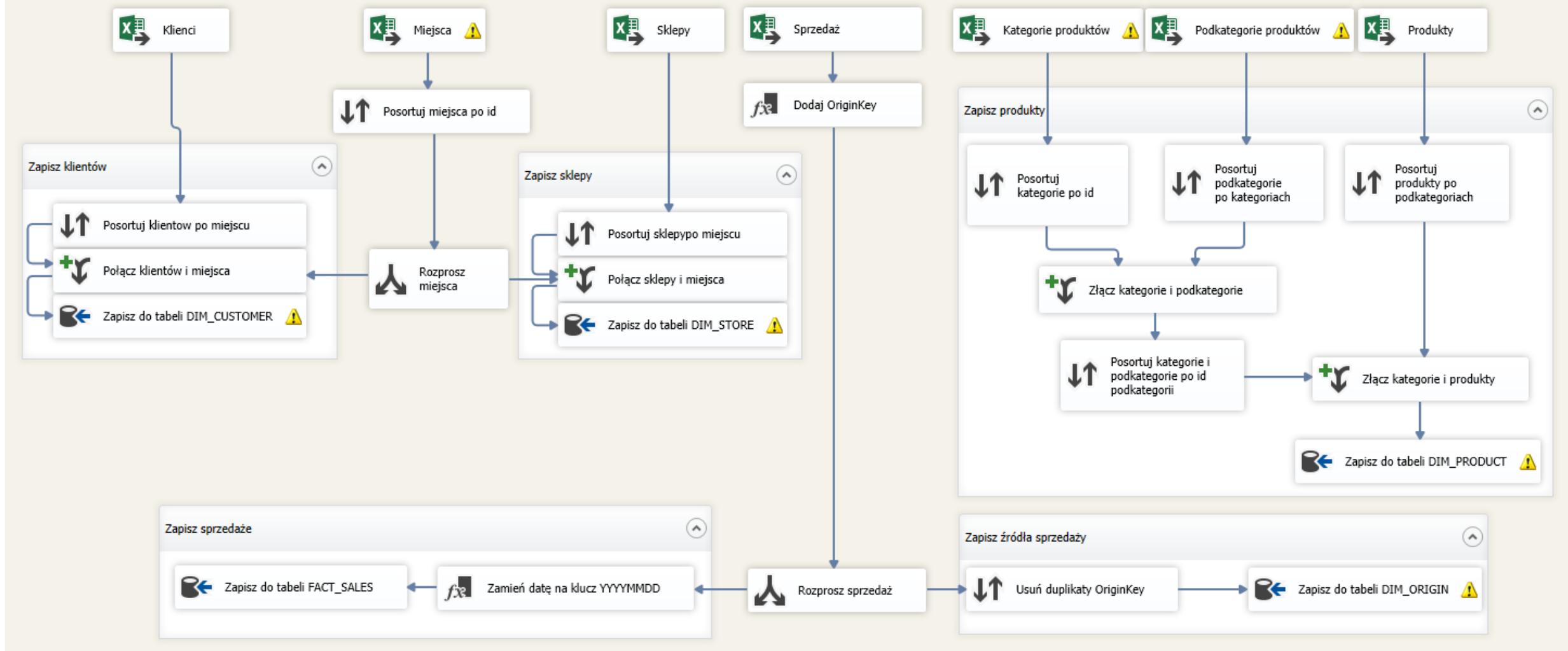


„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”

# Przykłady procesów ETL



**„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”**





Fundusze  
Europejskie  
Wiedza Edukacja Rozwój



Politechnika Wrocławskiego



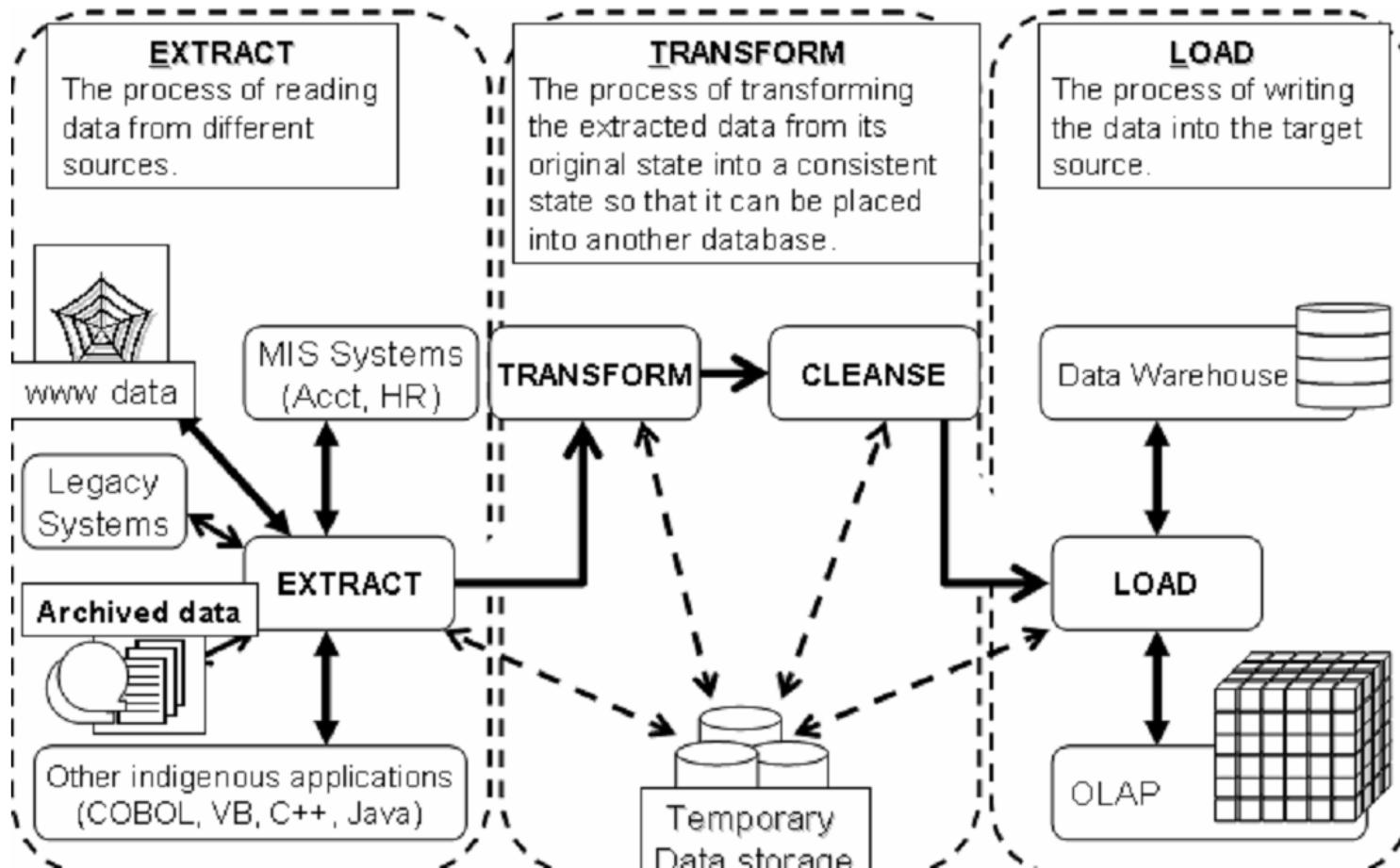
Unia Europejska  
Europejski Fundusz Społeczny

### „ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”





# ETL





Fundusze  
Europejskie  
Wiedza Edukacja Rozwój



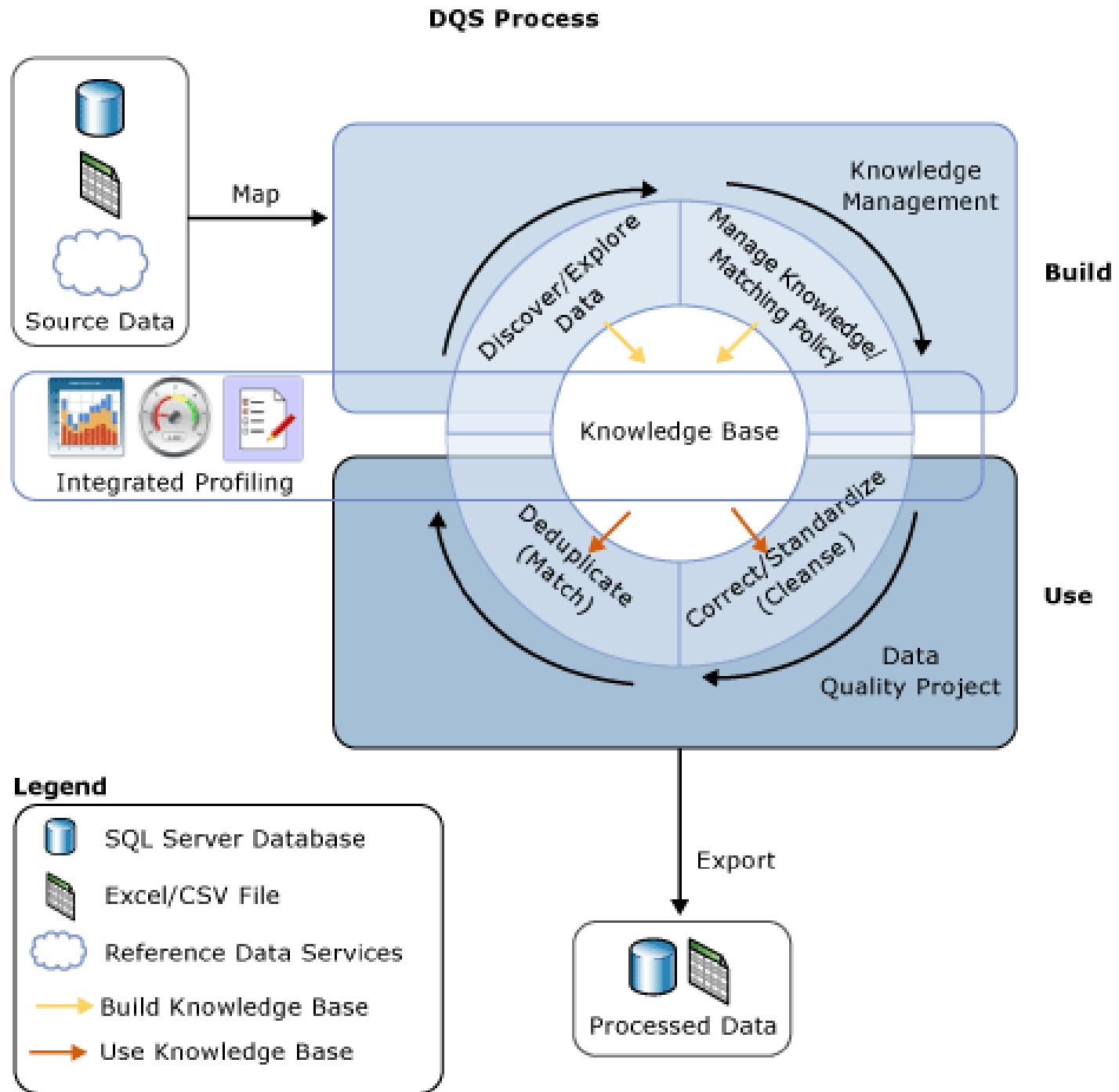
Politechnika Wrocławska

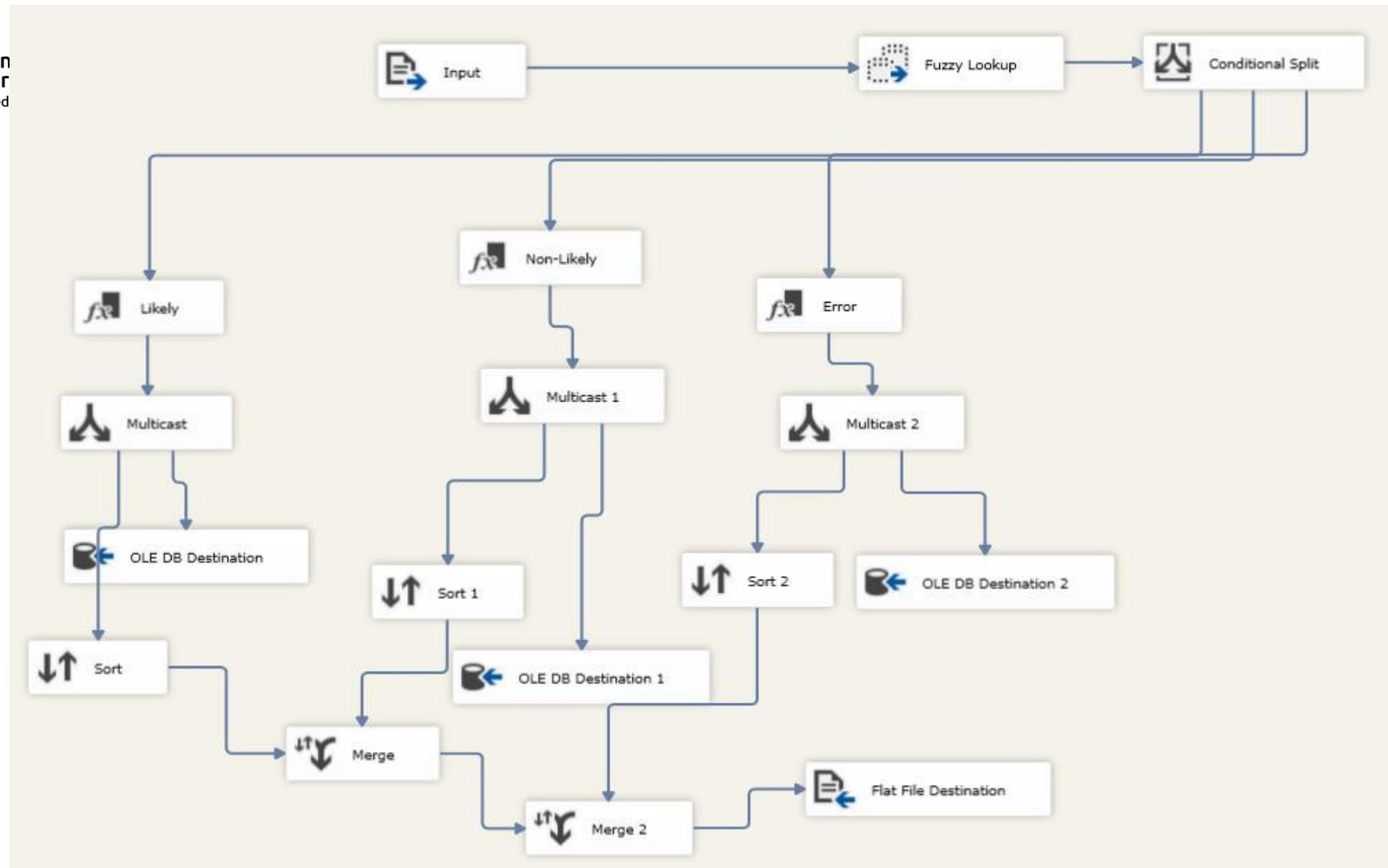
Unia Europejska  
Europejski Fundusz Społeczny



*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Data Quality Services





Error	Package	String	ERROR
likely	Package	String	LIKELY
non_match	Package	String	NON-MATCH

94	"Hex Nut 20	Hex Nut 20	0.9875	0.9007731	0.9875	LIKELY	opejska Spoteczny	
95	"HL Touring Seat/Saddle	HL Touring Seat/Sad...	0.9875	0.6114928	0.9875	LIKELY		
96	"Lock Washer 2	Lock Washer 2	0.9875	0.9578876	0.9875	LIKELY		
97	"Hex Nut 21	Hex Nut 21	0.9875	0.9007731	0.9875	LIKELY	j"	
98	"LL Bottom Bracket	LL Bottom Bracket	0.9875	0.5996314	0.9875	LIKELY		
99	"HL Mountain Rim	HL Mountain Rim	0.9875	0.681129	0.9875	LIKELY		
100	"Hex Nut 2	Hex Nut 2	0.9875	0.9713477	0.9875	LIKELY		
101	"Lock Washer 11	Lock Washe	"Bearing Ball,Bearing Ball,0.98750001,0.5,0.98750001,LIKELY					
102	"Lock Washer 5	Lock Washe	"External Lck Washer 8,External Lock Washer 8,0.92961943,0.56844395,0.92961943,NON-MATCH					
103	"Thin-Jam Lock Nut 13	Thin-Jam Loc	"External Lock Washer 1,External Lock Washer 1,0.98750001,0.98534936,0.98750001,LIKELY					
104	"Lock Washar 3	Lock Washar	"External Lock Washer 7,External Lock Washer 7,0.98750001,0.56334531,0.98750001,LIKELY					
105	"ML Grip Tpe	ML Grip Tap	"External Lock Washer 9,External Lock Washer 9,0.98750001,0.56254739,0.98750001,LIKELY					
106	"External Lck Washer 8	External Locl	"Guide Pulley,Guide Pulley,0.98750001,0.56722081,0.98750001,LIKELY					
			"Headset Ball Bearings,Headset Ball Bearings,0.98750001,0.52441591,0.98750001,LIKELY					
			"Hex Nut 1,Hex Nut 1,0.98750001,0.97134769,0.98750001,LIKELY					
			"Hex Nut 10,Hex Nut 10,0.98750001,0.94661856,0.98750001,LIKELY					
			"Hex Nut 11,Hex Nut 11,0.98750001,0.93934381,0.98750001,LIKELY					
			"Hex Nut 12,Hex Nut 12,0.98750001,0.93934381,0.98750001,LIKELY					
			"Hex Nut 13,Hex Nut 13,0.98750001,0.93934381,0.98750001,LIKELY					
			"Hex Nut 16,Hex Nut 16,0.98750001,0.93017125,0.98750001,LIKELY					
			"Hex Nut 17,Hex Nut 17,0.98750001,0.90077311,0.98750001,LIKELY					
			"Hex Nut 2,Hex Nut 2,0.98750001,0.97134769,0.98750001,LIKELY					
			"Hex Nut 20,Hex Nut 20,0.98750001,0.90077311,0.98750001,LIKELY					
			"Hex Nut 21,Hex Nut 21,0.98750001,0.90077311,0.98750001,LIKELY					
			"Hex Nut 22,Hex Nut 22,0.98750001,0.90077311,0.98750001,LIKELY					
			"Hex Nut 23,Hex Nut 23,0.98750001,0.90077311,0.98750001,LIKELY					
			"Hex Nut 3,Hex Nut 3,0.98750001,0.9654057,0.98750001,LIKELY					
			"Hex Nut 5,Hex Nut 5,0.98750001,0.96176714,0.98750001,LIKELY					
			"Hex Nut 7,Hex Nut 7,0.98750001,0.96176714,0.98750001,LIKELY					
			"Hex Nut 8,Hex Nut 8,0.98750001,0.96176714,0.98750001,LIKELY					
			"Hex Nut 9,Hex Nut 9,0.98750001,0.95754081,0.98750001,LIKELY					



„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”

# Fuzzy Grouping

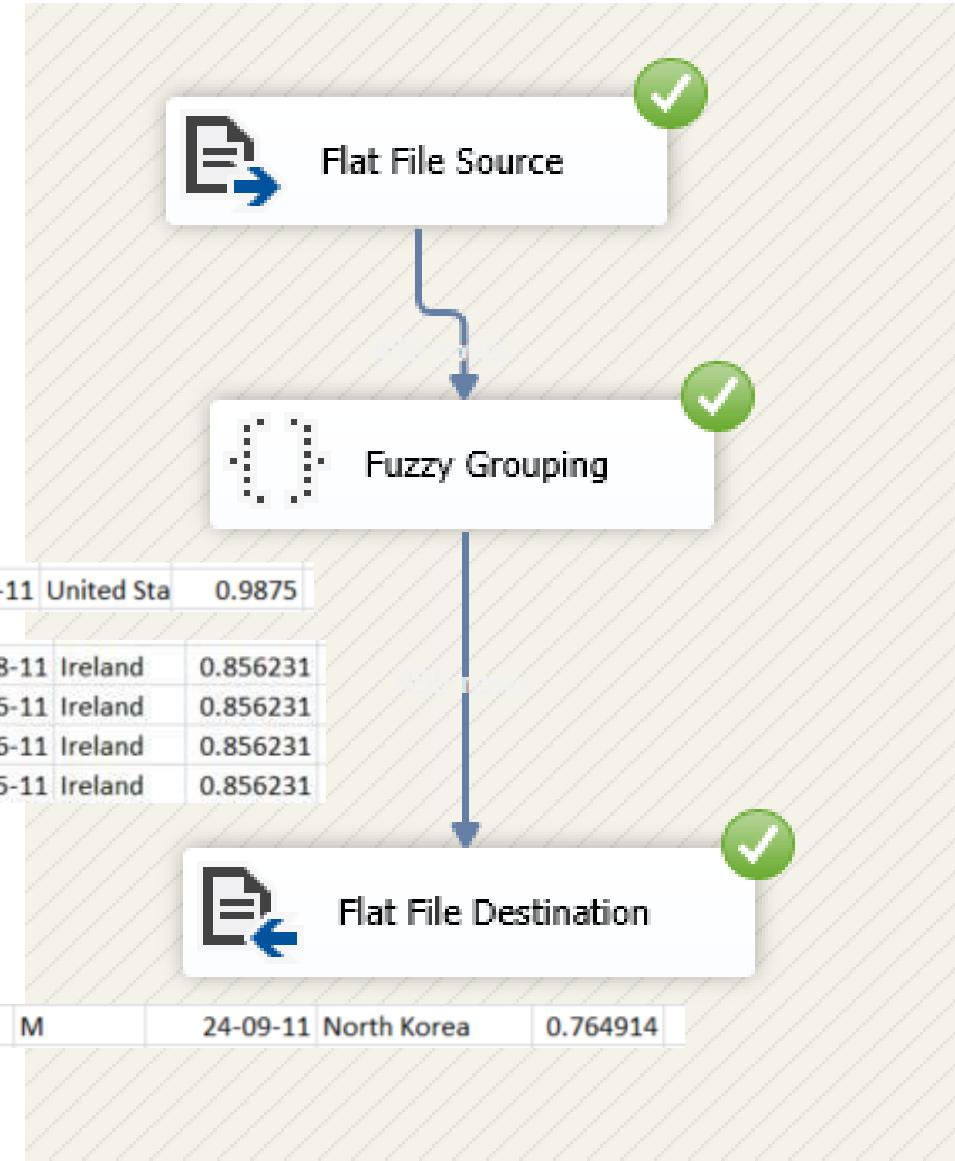
The screenshot shows a list of columns with checkboxes:

City	<input checked="" type="checkbox"/>
Continent	<input type="checkbox"/>
ReviewerName	<input checked="" type="checkbox"/>
Gender	<input checked="" type="checkbox"/>
Time	<input checked="" type="checkbox"/>

Below the list is a configuration table:

Input Column	Output Alias	Group Output Alias	Match Type	Minimum Similarity	Similarity
City	City	City_clean	Fuzzy	0	_Simila

486	246	0.9875	490	346	0.824277	United-States	South America	Sherman	M	31-10-11	United States	0.9875
326	366	0.856231	330	981	2.022023	Iceland	Africa	Bryan	Phi F	31-08-11	Ireland	0.856231
177	366	0.856231	181	534	10	Iceland	Europe	Sherri	Ho M	28-06-11	Ireland	0.856231
113	366	0.856231	117	768	4.09318	Iceland	South America	Guillermo	F	09-06-11	Ireland	0.856231
21	366	0.856231	25	713	5.871027	Iceland	South America	Geraldine	F	07-05-11	Ireland	0.856231



377	187	0.764914	381	792	2.603801	South Korea	South America	Kristi	Brown	M	24-09-11	North Korea	0.764914
-----	-----	----------	-----	-----	----------	-------------	---------------	--------	-------	---	----------	-------------	----------

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

## Tworzenie partycji

- fizyczny podział tabeli faktów na mniejsze tabele
- cel: poprawa wydajności zapytań
- zazwyczaj podział względem dat
- uwzględnienie tych części wymiarów, które są potrzebne

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Dziennik zmian

- nadmiarowe, zbędne
- wszystkie dane są wprowadzane procesem ETL
- dane ładowane są luzem
- w przypadku niepowodzenia, proces można powtórzyć
- różne systemy bazodanowe korzystają z różnych dzienników
  - jak je zintegrować?



Fundusze  
Europejskie  
Wiedza Edukacja Rozwój



Politechnika Wrocławska

Unia Europejska  
Europejski Fundusz Społeczny



*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Hurtownie danych

Dziękuję za uwagę

dr inż. Marcin Maleszka



Fundusze  
Europejskie  
Wiedza Edukacja Rozwój



Politechnika Wrocławskiego

Unia Europejska  
Europejski Fundusz Społeczny



*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Hurtownie danych

**Logiczna organizacja hurtowni danych**

**dr inż. Marcin Maleszka**



*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Model konceptualny

- Warstwa pojęciowo - funkcjonalna
  - operuje na poziomie informacji, definiuje funkcje (biznesowe) HD
  - posługuje się językiem pojęć biznesowych
- Warstwa logiczna
  - operuje na poziomie informacji i danych
  - mapuje pojęcia biznesowe (informację) na język danych
- Warstwa fizyczna
  - operuje na poziomie danych
  - stanowi implementację warstwy logicznej

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

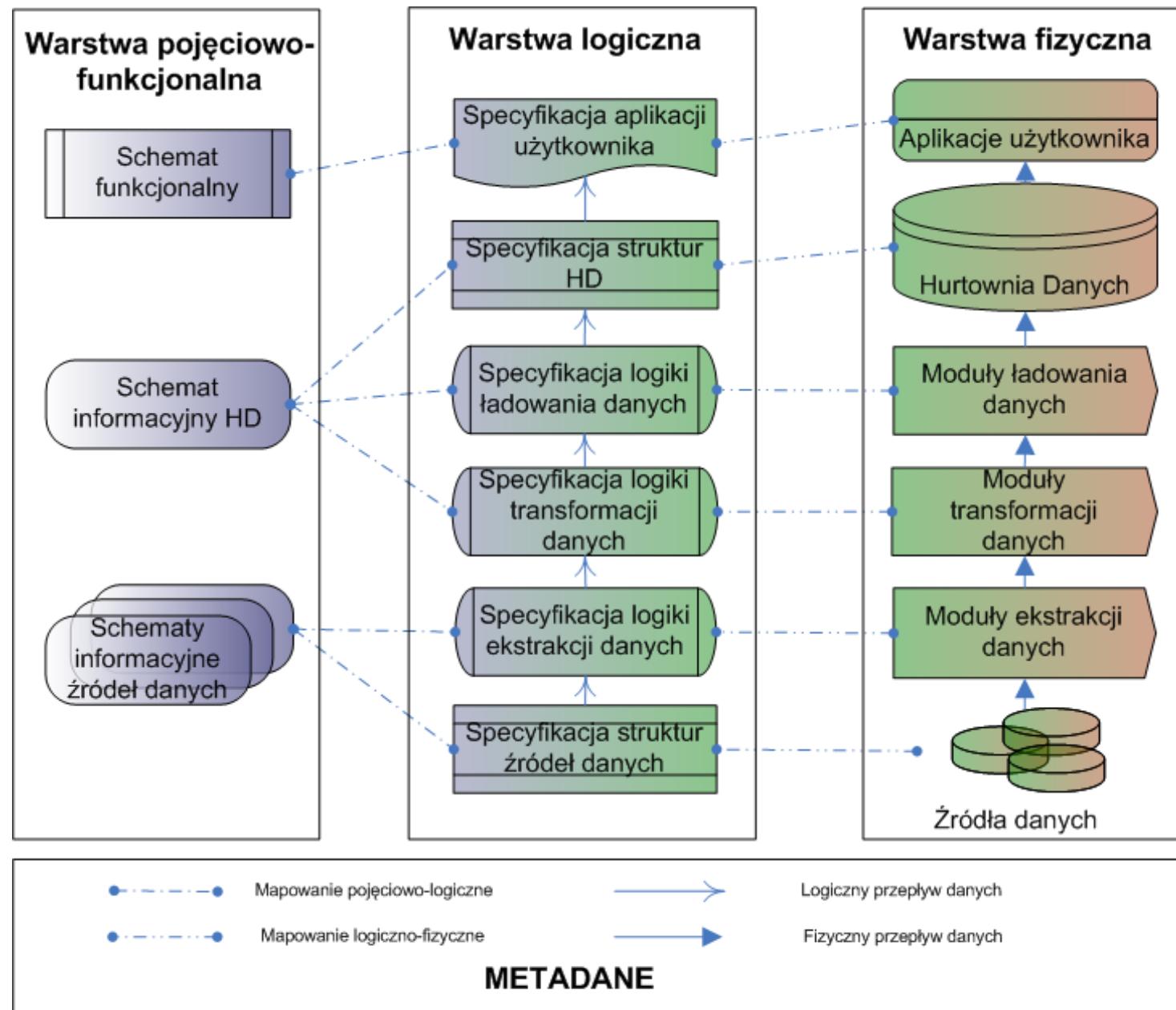
# Modelowanie hurtowni danych

- Model biznesowy
  - Efekt analizy strategicznej
  - Identyfikacja miar i wymiarów dla poszczególnych procesów biznesowych
- Model logiczny (wymiarowy)
  - Model abstrakcyjny, konceptualny
  - Encje i atrybuty (reprezentowane w modelu relacyjnym jako tabele i powiązania między nimi)

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Modelowanie hurtowni danych

- Model fizyczny
  - Wybór sposobu składowania danych
  - Formaty danych
  - Strategie partycjonowania
  - Wybór indeksów
  - Wybór materializowanych perspektyw



*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Warstwa logiczna

- Specyfikacja aplikacji użytkownika
- Specyfikacja struktur HD
- Specyfikacja struktur ETL
  - Specyfikacja logiki ekstrakcji danych
  - Specyfikacja logiki transformacji danych
  - Specyfikacja logiki ładowania danych
- Specyfikacja struktur źródeł danych



*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Model struktury danych

- Fakty (zdarzenia)
  - ujęcie ilościowe miary – np. cena, liczba sztuk, wartość, itp.
  - główna tabela HD
- Wymiary
  - nadają kontekst faktom – np. kto, gdzie, kiedy, jak, itp..
  - odrębne tabele
  - mogą być hierarchiczne
- Projekt logiczny HD: określenie faktów i wymiarów je opisujących

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

## Rodzaje wymiarów

- Wymiar uzgodniony (ang. conformed dimension)
  - wymiar, który pozostając w relacji z wieloma faktami ma takie samo znaczenie
  - Skąd problem?
  - Integracja różnych źródeł
  - Dwa wymiary są uzgodnione jeśli są identyczne lub jeden jest podzbiorem drugiego
- Wymiar wielokrotnego stosowania (ang. role-playing dimension)
  - Wymiar przechowywany w jednej tabeli, ale wykorzystywany wielokrotnie, np. data

## Rodzaje wymiarów

- Wymiar abstrakcyjny (ang. junk dimension)
  - Połączenie różnych wymiarów o małej liczebności atrybutów, np. płeć i grupa wiekowa
  - Iloczyn kartezjański „małych” wymiarów
- Wymiar zdegenerowany (ang. degenerate dimension)
  - Liczność wymiaru jest porównywalna z liczbą faktów
  - Klucz biznesowy przechowywany w tabeli faktów

# Wolno zmieniające się wymiary

- Przyczyny:
  - Powiązanie pozycji wymiaru z faktem jest zmieniane lub anulowane
  - Wartości atrybutów pozycji wymiaru ulegają zmianie (w kontekście czasu) w rozpatrywanym wycinku rzeczywistości
- Typy:
  - Zmiana traktowana jest jako błąd (Typ 0)
  - Pamiętana jest ostatnia wartość (nadpisanie -Typ 1)
  - Pamiętana jest cała historia zmian (Typ 2)
  - Pozostawia się historię zmian w ograniczonym zakresie np. trzy ostatnie zmiany (Typ 3)
  - Nieaktualne dane przenieś do tabeli historycznej (Typ 4)

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Który to typ?

Pracownik_ID	PESEL	Imię	Nazwisko	Stanowisko
004352	90120923877	Katarzyna	Nowak	Sprzedawca

Pracownik_ID	PESEL	Imię	Nazwisko	Stanowisko
004352	90120923877	Katarzyna	Kowalska	Sprzedawca



Fundusze  
Europejskie  
Wiedza Edukacja Rozwój



Politechnika Wrocławskiego

Unia Europejska  
Europejski Fundusz Społeczny



„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”

## Który to typ?

Pracownik_ID	PESEL	Imię	Nazwisko	Stanowisko
004352	90120923877	Katarzyna	Nowak	Sprzedawca

Praconik_ID	PESEL	Imię	Nazwisko	Stanowisko	Od	Do	Status
004352	90120923877	Katarzyna	Nowak	Sprzedawca	2014-07-01	2019-03-26	nieaktualne
0049872	90120923877	Katarzyna	Nowak	Kierownik	2019-03-27	9999-12-31	aktualne



Fundusze  
Europejskie  
Wiedza Edukacja Rozwój



Politechnika Wrocławskiego

Unia Europejska  
Europejski Fundusz Społeczny



„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”

## Który to typ?

Pracownik_ID	PESEL	Imię	Nazwisko	Stanowisko
004352	90120923877	Katarzyna	Nowak	Sprzedawca

Praconik_ID	PESEL	Imię	Nazwisko	Aktualne stanowisko	Poprzednie stanowisko
004352	90120923877	Katarzyna	Nowak	Kierownik	Sprzedawca

# Wymiary szybkozmienne

- Atrybut lub grupa atrybutów zmienia się szybko, ale w ograniczonym zakresie
  - szybkość definiowana jest przez rzeczywistość biznesową

Rozwiązania:

- Miniwymiar
  - tabela odpowiadająca wszystkim dopuszczalnym wartościom
  - jeśli zmienia się kilka atrybutów, to łączymy te miniwymiary w wymiar abstrakcyjny
- Dodatkowa tabela faktów typu „fakty bez faktów”
  - łączy wymiar (atrybuty nieszybkozmienne) i miniwymiar lub wymiar abstrakcyjny
  - może łączyć się też z innym wymiarem, np. czas wprowadzenia zmiany

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Wymiar hierarchiczny

- Hierarchie występujące naturalnie
- Zbalansowanie hierarchii
- Hierarchia typu rodzic-dziecko
- Domyślny element wymiaru

## Tabela faktów - rodzaje

- Fakty bez faktów (ang. factless facts)
  - Fakty bez miar, samo zdarzenie jest faktem
  - Przykłady: wizyta pacjenta u lekarza, strzelenie bramki przez zawodnika, udział poszkodowanego w kolizji/wypadku, itp. Możliwa jest sytuacja, że mamy informacje o miarach przy tych zdarzeniach.
- Brakujące miary
  - NULL – niezbędne zabezpieczenie IF NOT IsNull (atrribut)
  - Wartość domyślna – problem ze znaczeniem
  - Wartość domyślna + dodatkowa flaga boolowska ((true, 1), (false, 0))

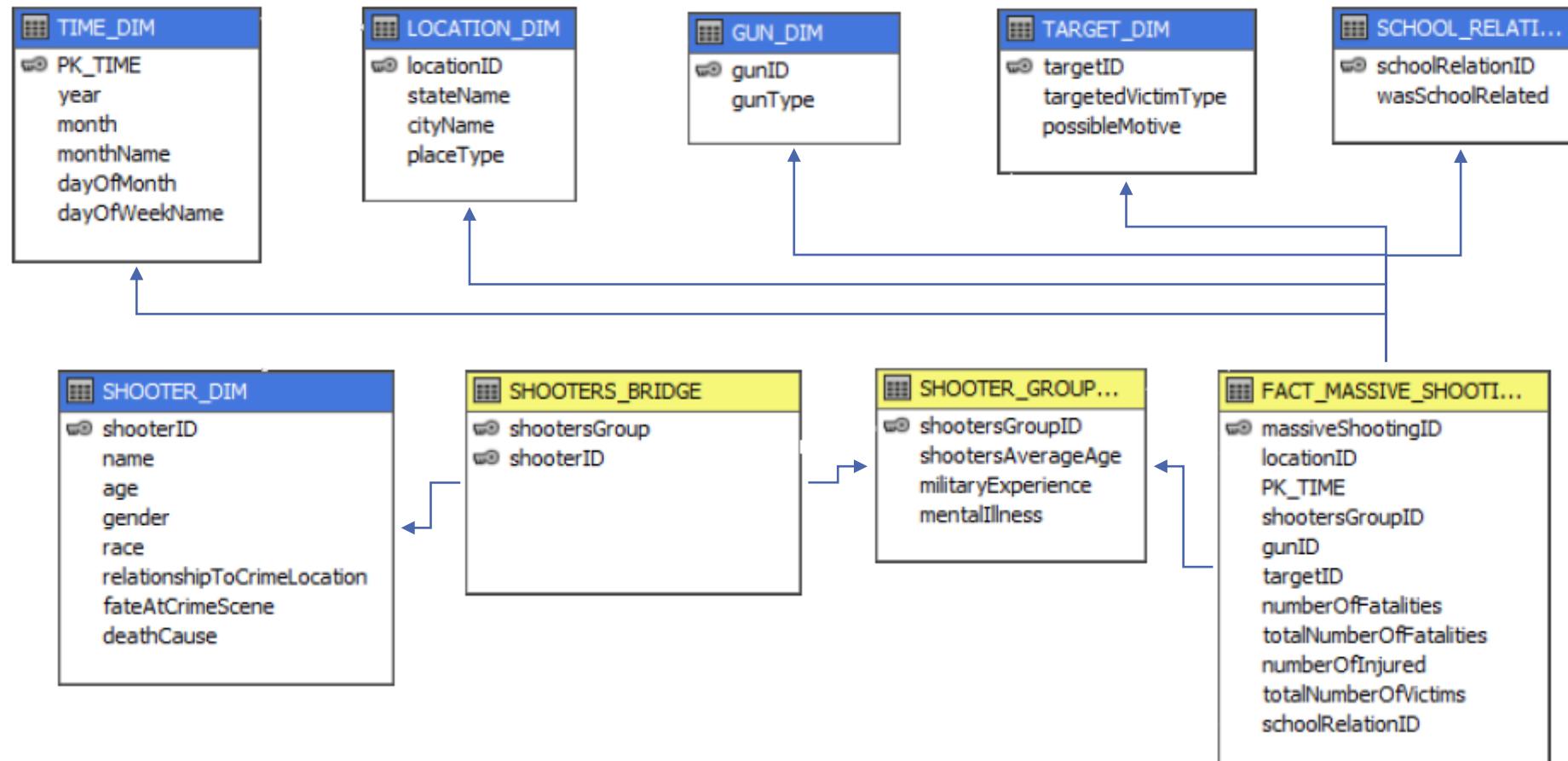
## Most (ang. bridge)

- W efekcie denormalizacji możemy uzyskać tabele połączone relacją wiele-do-wielu
  - wielowartościowe wymiary, np. wielu autorów utworu muzycznego
  - wielowartościowe atrybuty, np. wiele umiejętności/zainteresowań pracownika
- Pytania:
  - jak wyliczyć zysk ze sprzedaży konkretnego albumu / sumaryczny zysk pojedynczego autora?
  - jak przygotować zestawienie zysku firmy w zależności od elementarnych umiejętności pracownika?



„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”

# Most



# Sposoby uniknięcia mostów

- Zmiana ziarnistości faktów
  - przykład: wiersz to transakcja -> wiersz to konkretny produkt z transakcji
- Wskazanie wartości „pierwotnej”
  - wskazanie podstawowej wartości z jednym kluczem obcym
  - oznaczenie odpowiedniego atrybutu jako „pierwotny”/”podstawowy”
- Dodanie wielu atrybutów do tabeli wymiarów
  - przykład: kolejne atrybuty to nazwy umiejętności -> wartości atrybutów true/false
  - skalowalność -> wystarczające dla stałej, ograniczonej liczby wartości

# Sposoby uniknięcia mostów

- Dodanie kolumny zawierającej konkatenację wartości atrybutów do wymiaru
  - możliwe tylko dla ograniczonej liczby wartości
  - niezbędny ogranicznik pomiędzy kolejnymi wartościami
  - łatwa prezentacja
  - trudność z zapytaniami (wieloznaczne wyszukiwania, wolniejsze działanie)
  - problemy z wyliczaniem sum miar
  - problemy z grupowaniem/filtrowaniem według konkretnych wartości atrybutów

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Projektowanie miar

- Jakie dane mogę zastosować jako miary?
- Jakich miar wyliczanych potrzebuję?
- Czy jest zapewniona addytywność miar?

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Specyfikacja struktur ETL

- ETL – ogół działań mających na celu:
  - pobranie danych ze źródeł,
  - przekształcenie ich do postaci pożądanej w modelu HD
  - sprawdzenie ich poprawności
  - wykonanie niezbędnych operacji, np. agregacji
  - załadowanie danych do struktur HD

# Specyfikacja logiczna ETL

- Extract
  - określenie typu i zakresu danych źródłowych
  - określenie sposobu pozyskania danych
  - określenie warunków dostępności i form transmisji danych
  - zapewnienie systemu kontroli poprawności ekstrakcji
- Transform:
  - przekształcenie typów i wartości
  - przekształcenie do pożądanej struktury
  - zapewnienie poprawności i spójności ekstraktów
  - zapewnienie wydajności procesu przetwarzania, np. uwzględnienie właściwej kolejności operacji



Fundusze  
Europejskie  
Wiedza Edukacja Rozwój



Politechnika Wrocławskiego

Unia Europejska  
Europejski Fundusz Społeczny



*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Specyfikacja logiczna ETL

- Load:
  - wczytanie poprawnych danych w odpowiedniej kolejności
  - mapowanie elementów do odpowiednich struktur HD
  - zasilanie tematycznych HD
  - wyliczanie predefiniowanych raportów i analiz

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Podsumowanie

- Warstwa logiczna odpowiada projektowi technicznemu w inżynierii oprogramowania
- Specyfikacja powinna zawierać opis wszystkich algorytmów
- Warstwa logiczna jest niezbędna do rozpoczęcia implementacji



Fundusze  
Europejskie  
Wiedza Edukacja Rozwój



Politechnika Wrocławska

Unia Europejska  
Europejski Fundusz Społeczny



*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Hurtownie danych

Dziękuję za uwagę

dr inż. Marcin Maleszka



Fundusze  
Europejskie  
Wiedza Edukacja Rozwój



Politechnika Wrocławska

Unia Europejska  
Europejski Fundusz Społeczny



*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Hurtownie danych

**Architektura hurtowni danych**

**dr inż. Marcin Maleszka**

# Architektura

- Ogólnie:
- Zbiór zasad i struktur będących szkieletem ogólnego projektu systemu lub produktu
- Struktura hurtowni danych - kolejne warstwy danych, przy czym każda następna warstwa stanowi przetworzenie poprzedniej:
  - **źródła danych**, czyli zastane bazy danych – rozproszone, niejednorodne, często niespójne
  - centralna hurtownia danych – podstawowe miejsce przechowywania nieulotnej informacji
  - hurtownie tematyczne, lokalne
  - aplikacje użytkowników
  - dodatkowo: baza metadanych

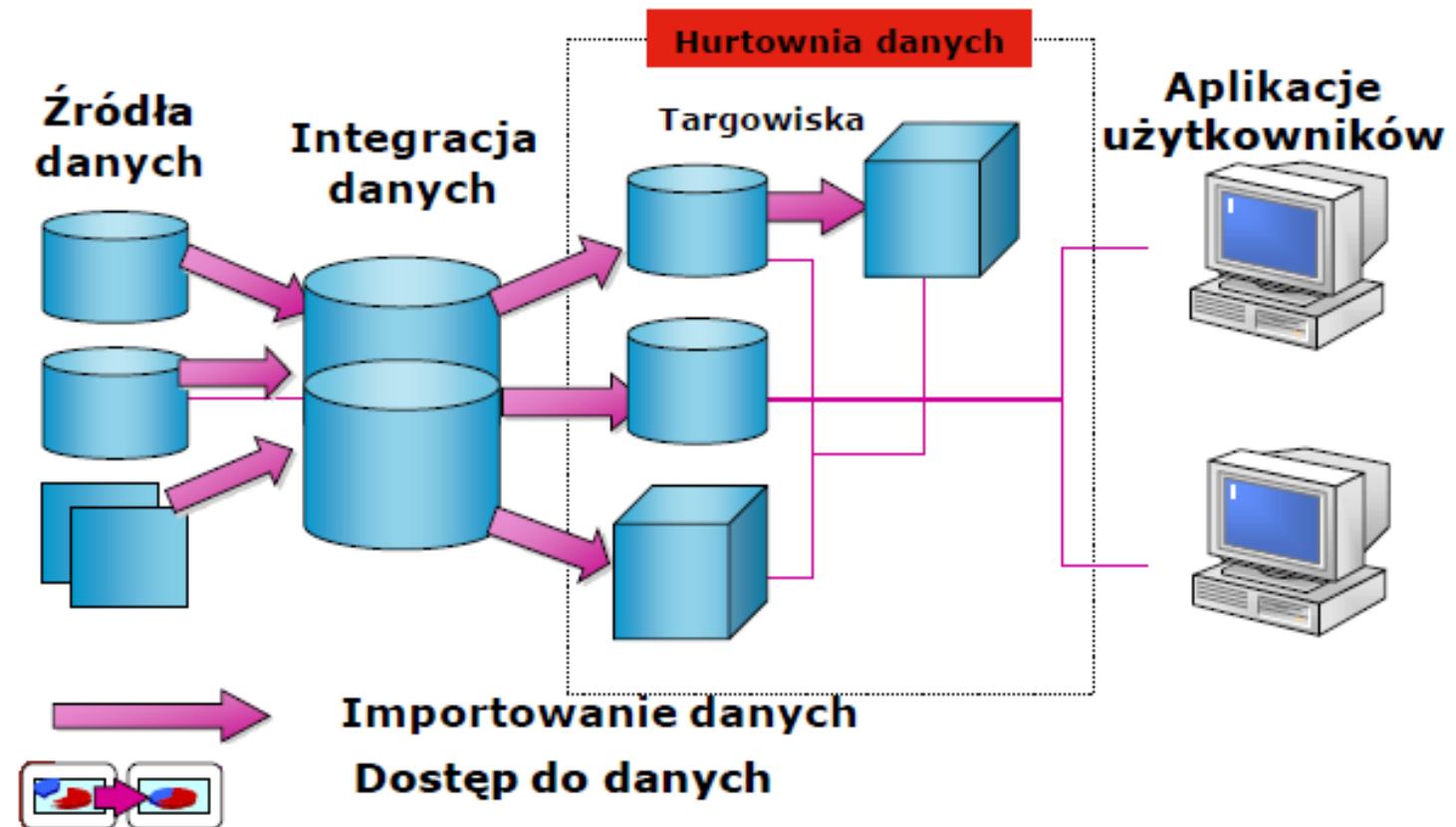


*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Cechy architektury danych dla hurtowni danych

- dane pochodzą z systemów źródłowych, baz danych lub plików
- dane z systemów źródłowych podlegają procesom ETL przed wprowadzeniem ich do hurtowni
- trwała analityczna baza danych jest przystosowana do przetwarzania wspomagającego podejmowanie decyzji
- użytkownicy mają dostęp do hurtowni przez specjalnie zaprojektowane narzędzia

# Hurtownia danych



# Rozwinięcia architektury ogólnej

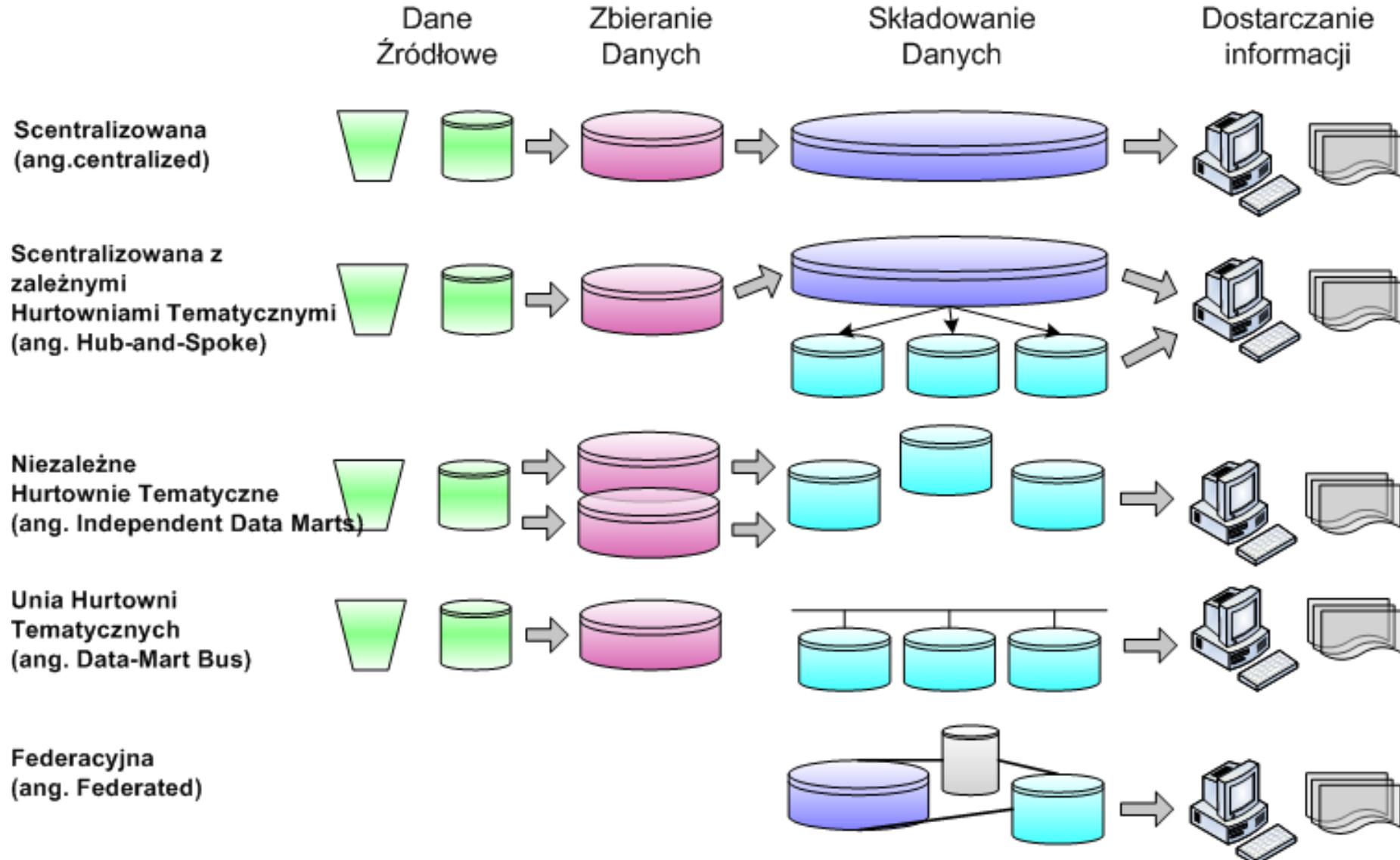
- Konkretnie wdrożenia rozwijają poszczególne komponenty architektury ogólnej
- Zwykle dąży się do integracji architektury HD z architekturą aktualnie działającego systemu
- Przykłady rozwinięć:
  - wyodrębnienie hurtowni tematycznych jako element pośredni pomiędzy centralną hurtownią a aplikacjami użytkownika
  - hurtownia tematyczna zbudowana bezpośrednio po integracji danych
  - zintegrowana relacyjna baza danych jako element pośredni pomiędzy fazą integracji danych a HD

# Typowe modele hurtowni danych

- Architektura zcentralizowana
  - materialna centralna hurtownia danych
- Architektura federacyjna
  - wirtualna centralna hurtownia danych jako niezmaterializowana perspektywa o wspólnym schemacie logicznym i pojęciowym
  - dane fizycznie przechowywane w magazynach danych operacyjnych
  - spadek wydajności ze względu na możliwość rozproszenia magazynów danych operacyjnych
- Architektura warstwowa
  - wiele warstw hurtowni tematycznych zawierających coraz wyższe stopnie agregacji danych
  - wszystkie warstwy zmateriaлизowane (wydajność)
  - optymalizacja wielkości danych w hurtowniach tematycznych



*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*



*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Infrastruktura hurtowni danych

- Ścisłe powiązana z architekturą
- Wszelkie techniki, systemy, platformy, bazy danych, bramy i itp. niezbędne do zrealizowania wybranej architektury
- Cechy architektur BI:
  - skalowalność
  - dostępność
  - bezpieczeństwo
  - zarządzanie
  - współdziałanie

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Infrastruktura hurtowni danych

- Dane źródłowe:
  - format danych
  - sposób udostępnienia (dysk, sieć, protokół, itp.)
  - schemat zależności pomiędzy systemami
  - szkolenia dotyczące układów danych poszczególnych systemów źródłowych
- Integracja danych
  - narzędzia konwersji danych
  - narzędzia klasy ETL
  - sprzęt i oprogramowanie bazy danych
  - sieć i bramy łączące źródła danych z bazą

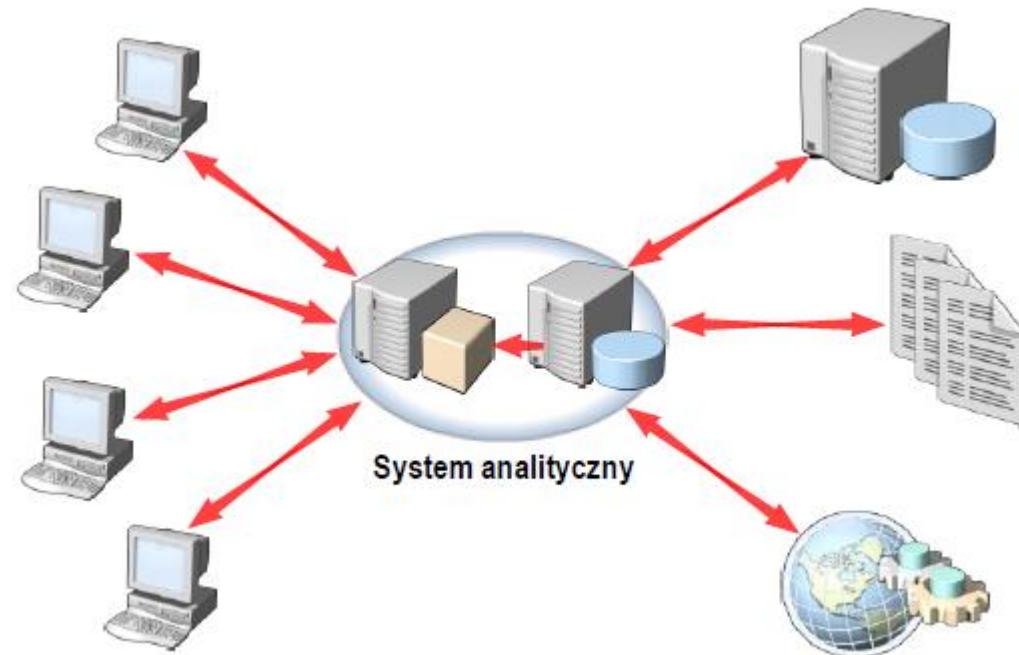
*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Infrastruktura hurtowni danych

- Hurtownia danych
  - sprzęt i oprogramowanie hurtowni
  - technologia przechowywania agregacji
  - harmonogramowanie zadań
  - kompetencje administratora
- Narzędzia użytkownika
  - szkolenia dla użytkowników
  - metadane
  - narzędzia nawigacji po metadanych
  - osobiste narzędzia analityczne i raportujące
  - repozytoria raportów

## Typy OLAP

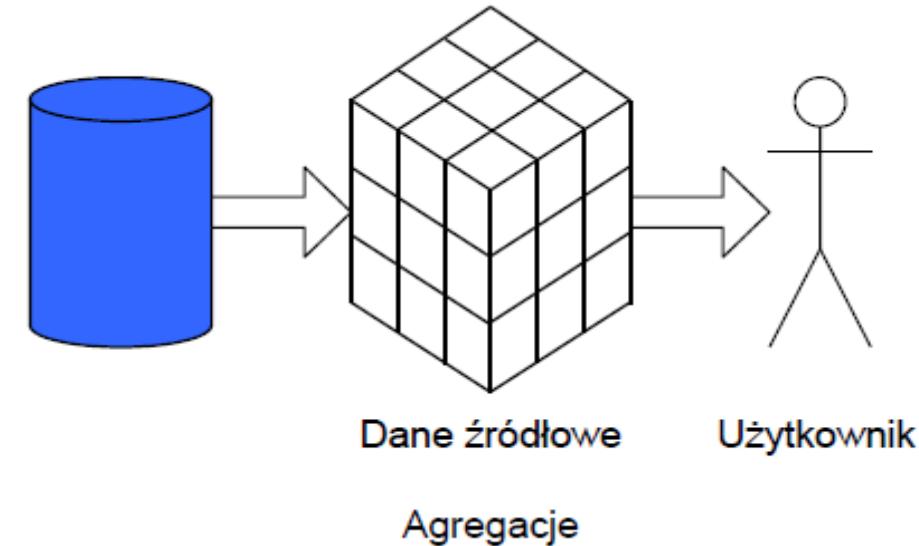
- bezpośrednie przetwarzanie analityczne
  - ROLAP
  - MOLAP
  - HOLAP



„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”

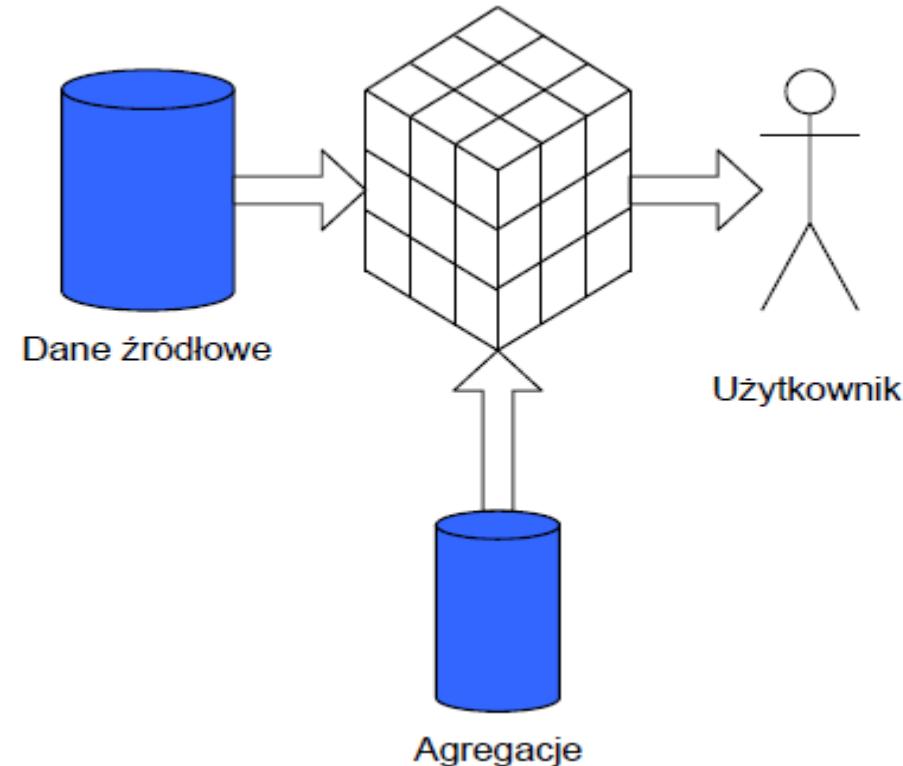
# MOLAP

- wielowymiarowe OLAP
- wielowymiarowy format danych i agregacji
- zapytania wykonują się szybko
- wymaga największej przestrzeni na dysku



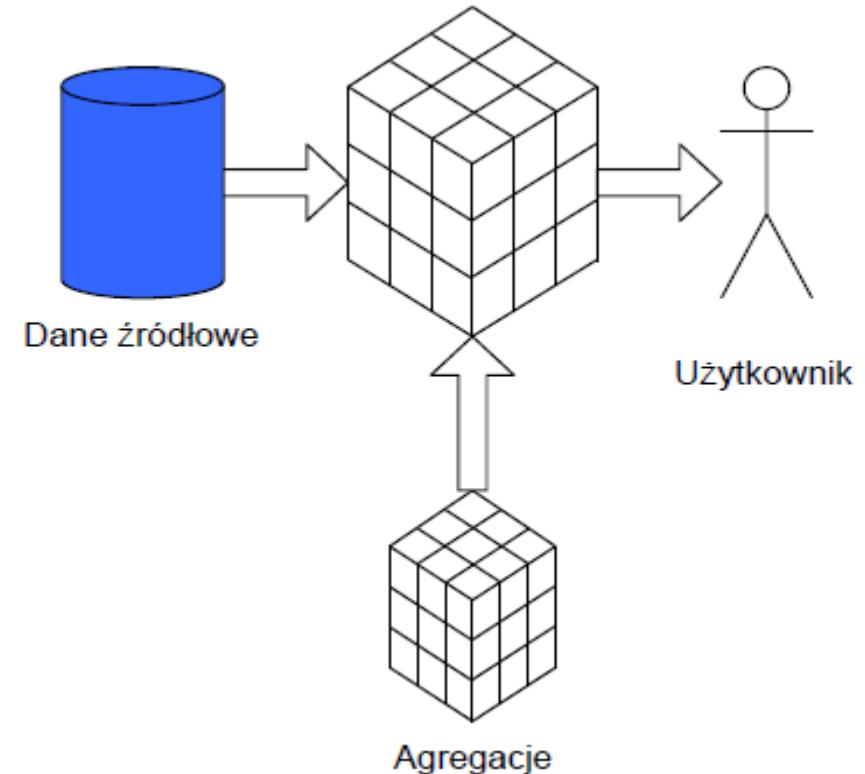
## ROLAP

- dane i agregacje są przechowywane w RSZBD
- najwolniejsze odpowiedzi na zapytania
- zwykle najwolniejsze przetwarzanie
- można tworzyć indeksowane widoki
- najbardziej użyteczny tryb dla dużej liczby danych
- wspomaga rozwiązania real-time OLAP



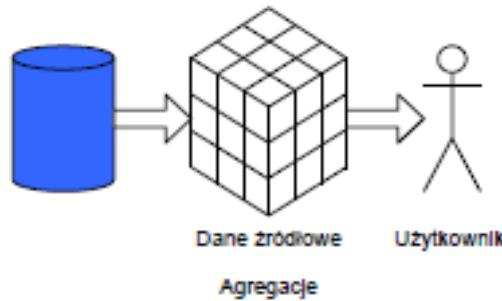
# HOLAP

- dane zarządzane przez RSZDB
- agregacje tworzone w formacie wielowymiarowym
- dobry wybór, gdy przestrzeń na dysku jest wąskim gardem
- dobra wydajność dla częstych odwołań do agregacji

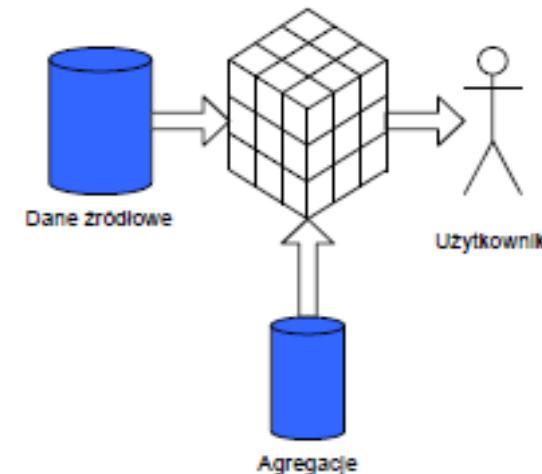


# Typy OLAP

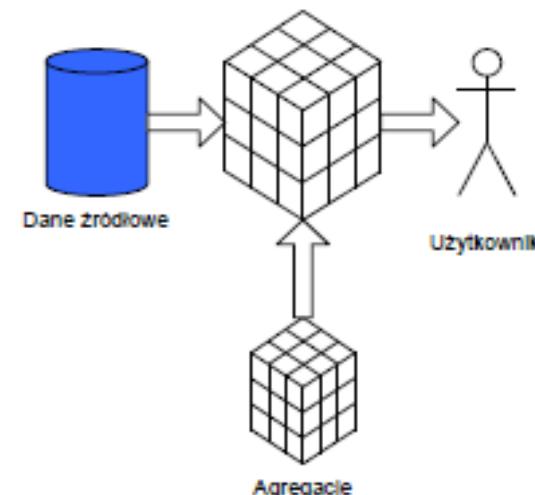
MOLAP



ROLAP



HOLAP



*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

## Typy systemów OLAP – podsumowanie

Model	opóźnienie	odpowiedzi	procesowanie	rozmiar
MOLAP	wysokie	szybko	szybko	średni
ROLAP	niskie	powoli	wolno	duży (?)
HOLAP	średnie	średnio	szybko	mały

# Konfiguracja i zarządzanie komponentami OLAP

- Zarządzanie operacjami w obszarze BI:
  - monitorowanie i optymalizacja rozwiązań BI
  - dokumentacja i procedura kontroli zmian
  - identyfikacja ról i narzędzi w obszarze BI
  - efektywne wsparcie dla rozwiązań BI



# Nadzorowanie zadań operacyjnych

- Nadzorowanie sporządzania kopii zapasowych
  - co zabezpieczać? kto wykonyuje kopie zapasowe?
  - harmonogram kopii zapasowych, czy kopie automatyczne?
  - walidacja kopii zapasowych
- Nadzorowanie i monitorowanie systemu zdarzeń systemowych
  - jakie zdarzenia nadzorować?
  - gdzie przechowywane są logi?
  - kto zarządza i sprawdza stan logów?
- Nadzorowanie procesowania kostek
  - określenie wymagań dotyczących procesowania i opóźnień

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Zarządzanie archiwizacją danych

- Planowanie archiwizacji
  - wymagania
  - efektywne przechowywanie danych
  - określenie wolumenów danych do archiwizacji
- Implementacja procesu archiwizacji
  - sposób realizacji archiwizacji
  - nośniki
  - miejsce przechowywania zarchiwizowanych danych
  - proces odzyskiwania lub odtwarzania awaryjnego

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Zachowanie ciągłości pracy

- Potrzeba ciągłości w przeprowadzaniu działalności operacyjnej
  - groźby wynikające z zachwiania ciągłości pracy
- Metody implementacji ciągłości pracy
  - system monitorowania narzędzi i procesów
  - procedury reakcji w przypadku zdarzeń systemowych
  - zasady zarządzania i odtwarzania w przypadku awarii
  - wymagania dotyczące zarządzania i przechowywania kopii zapasowych

# Konfiguracje dla dużych obciążień

- Rozdzielanie przetwarzania
  - odseparowane serwery dla każdej bazy OLAP
  - osobne serwery dla bazy źródłowej i bazy OLAP
  - osobny serwer dla procesowania
- Skalowanie dla dużych obciążień
  - kopie instancji na wielu serwerach
  - bilansowanie obciążzeń
  - procesowanie na jednym serwerze i synchronizacja baz OLAP

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Konfiguracje dla dużych obciążeń

- Opcje wysokiej dostępności
  - balansowanie obciążień, np. Network Load Balancing
  - klastry serwerów
- Problemy wysokiej dostępności:
  - sprzęt
  - system operacyjny
  - redundancja danych
  - dostępność kostek
  - przełączenie w przypadku awarii

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Zarządzanie dostępem

- Poziomy zabezpieczeń w systemie BI
  - rola i zabezpieczenia serwera
  - zabezpieczenie bazy danych
  - uprawnienia dostępu do obiektów
- Konfiguracja zabezpieczeń bazy OLAP
  - zabezpieczenia systemu operacyjnego
  - zabezpieczenia systemu plików
  - zabezpieczenie dostępu do źródła danych

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Zarządzanie dostępem

- Warunki dostępu administracyjnego
  - rola i zabezpieczenie serwera
  - uprawnienia administracyjne
  - członkostwo grup użytkowników
  - praca domenowa (grupowa)
- Warunki dostępu użytkowników
  - rola w bazie danych
  - uprawnienia wymiarów, kostek i komórek
  - filtry MDX



Fundusze  
Europejskie  
Wiedza Edukacja Rozwój



Politechnika Wrocławska

Unia Europejska  
Europejski Fundusz Społeczny



*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Hurtownie danych

Dziękuję za uwagę

dr inż. Marcin Maleszka



Fundusze  
Europejskie  
Wiedza Edukacja Rozwój



Politechnika Wrocławska

Unia Europejska  
Europejski Fundusz Społeczny



*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Hurtownie danych

**Podstawy MDX**

**dr inż. Marcin Maleszka**

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Wprowadzenie

- MDX jest Pseudo-akronimem dla **M**ulti**D**imensional **E**Xpressions
- Zastosowania:
- dla wielowymiarowych kalkulacji
- stosowany poza Analysis Services:
  - część specyfikacji OLE DB for OLAP
  - wielu innych dostawców używa MDX

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Zastosowania MDX

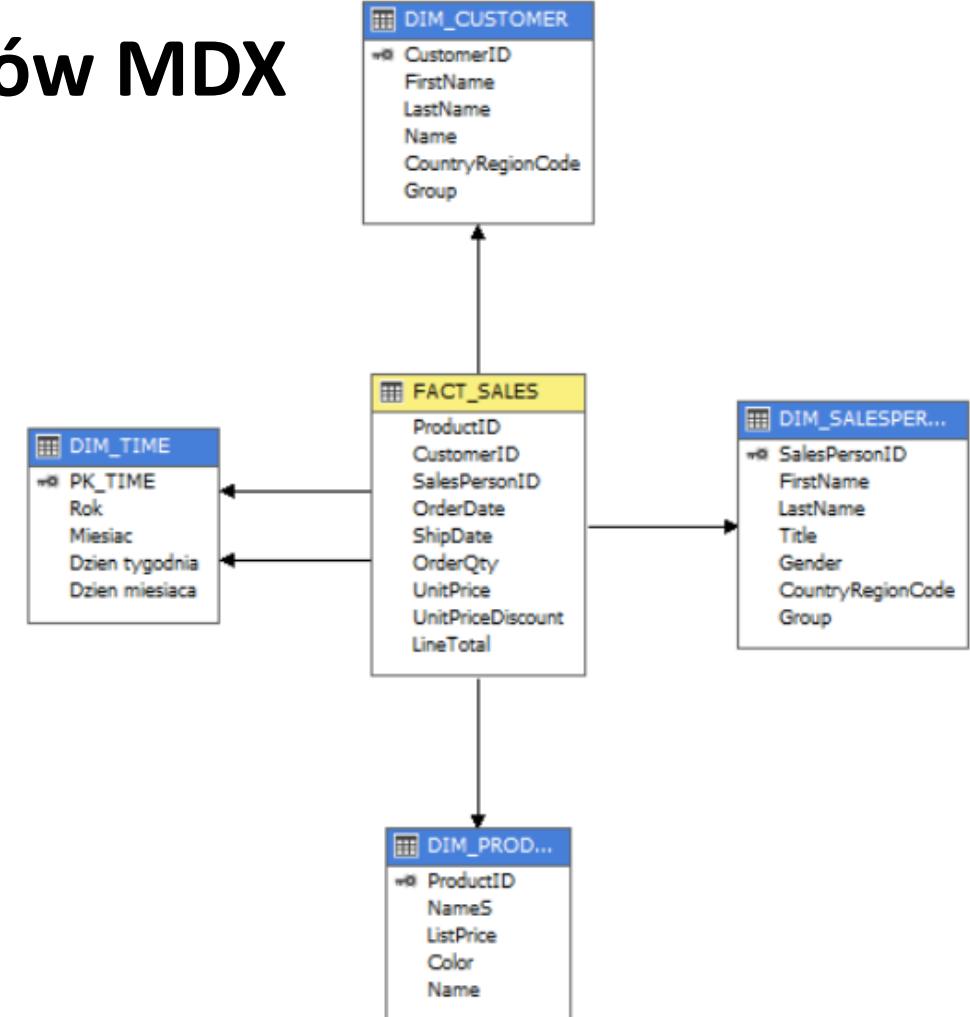
- Wyrażenie MDX
  - wielowymiarowa formuła podobna do funkcji w arkuszu kalkulacyjnym
  - używana do tworzenia elementów kalkulowanych, elementów domyślnych oraz własności wymiarów i kostek
  - rezultatem wykonania jest pojedyncza wartość
- Dyrektywa MDX
  - Język zapytań dla przeglądania zgromadzonych wyników podobny do zapytań SQL
  - Narzędzie dla aplikacji klienckich
  - Rezultatem wykonania jest kompletny zbiór wartości (np. tabela)

# Nazewnictwo obiektów MDX

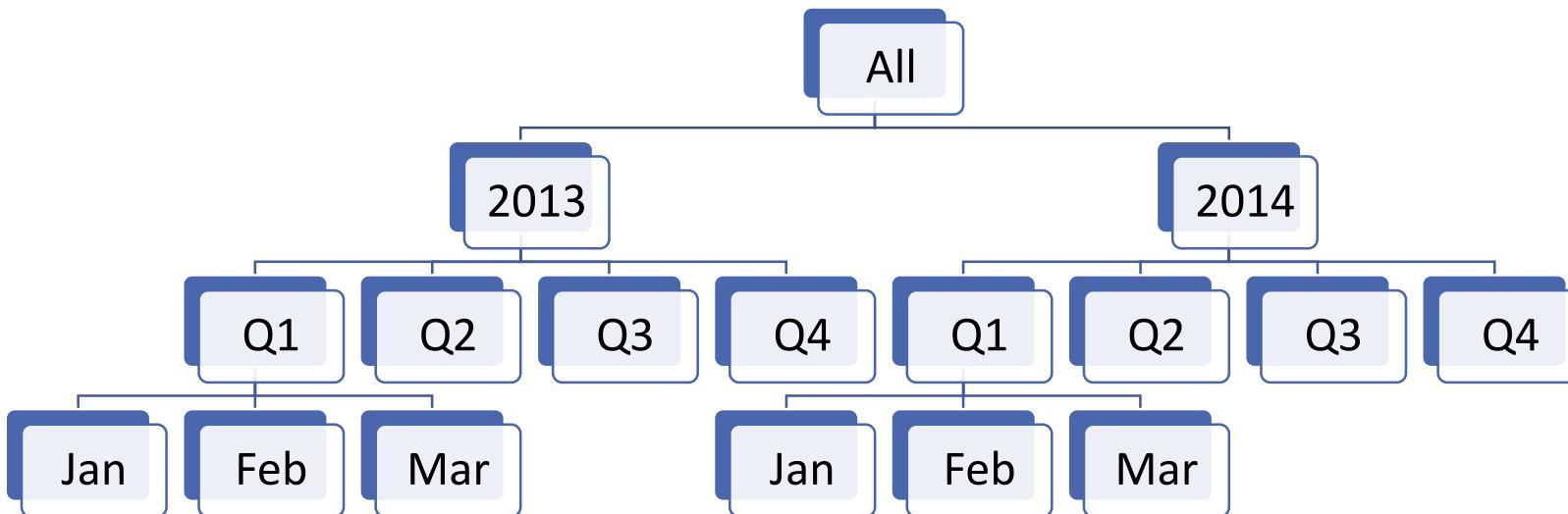
- Nazwa każdego obiektu OLAP musi być ujęta w nawiasy [ ] jeżeli:
  - zawiera spację lub innych znak specjalny, np. [Unit Price]
  - jeśli nazwa jest identyczna jak zastrzeżone słowo kluczowe lub funkcja, np. [Select]
  - jeśli nazwa rozpoczyna się znakami numerycznymi, np. [2016] lub [1200A]
- Kwalifikowane nazwy obiektów:
  - eliminacja niejasności:
    - [Q1] może być: [2014].[Q1] lub [2013].[Q1]
  - zasada „pierwszego wystąpienia”:
    - [Q1] niekwalifikowane - ?

# Nazewnictwo obiektów MDX

- Używamy nazwy elementu lub klucz:
  - [Country Region Code].[FR] lub [Country Region Code].&[3]
- Nazwa elementu jako:
  - element rodzica: [Country Region Code].[FR]
  - nazwa poziomu: [Group].[FR]
  - nazwa wymiaru: [Sales Person].[FR]
- Pełna nazwa kwalifikowana:
  - wymiar: [Sales Person].[Group].[Country Region Code].[FR]



## Wymiar: OrderDate



- Poprawne wywołania:

[OrderDate].[All].[2013].[Q1].[Feb]  
[2013].[Q1].[Feb]  
[(All)].[(All)].[2013].[Q1].[Feb]  
[Year].[2013].[Q1].[Feb]

- Problematyczne wywołania:

[OrderDate].[Feb]  
[2013].[Month].[Feb]  
[(All)].[2013].[Q1].[Feb]  
[Year].[2013].[Q2].[Feb]

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

## Cechy elementów kalkulowanych

- tworzone za pomocą wyrażeń MDX
- kalkulowane podczas przetwarzania zapytań, po agregacjach
- mogą się odwoływać do innych elementów lub miar, również kalkulowanych
- mogą należeć do dowolnego wymiaru
- nie zwiększają rozmiaru kostki i nie wymagają procesowania



„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”

# Wyrażenie MDX

**SELECT {wybrane atrybuty} on columns,  
{wybrane atrybuty} on rows**

**FROM NazwaKostki**

**WHERE (warunek)**

Numer osi	Alias
0	Columns
1	Rows
2	Pages
3	Section
4	Chapter

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Wyrażenia MDX

- Grają rolę funkcji
  - są kalkulowane dla każdej przeglądanej komórki
- jako rezultat zwracają wartości:
  - numeryczne
  - tekstowe
  - puste (ang. *Empty*, NULL)
- stosowane operatory:
  - dodawania (+), odejmowania (-), mnożenia (\*), dzielenia (/), potęgi (^)
- łączenie tekstów operatorem dodawania (+)

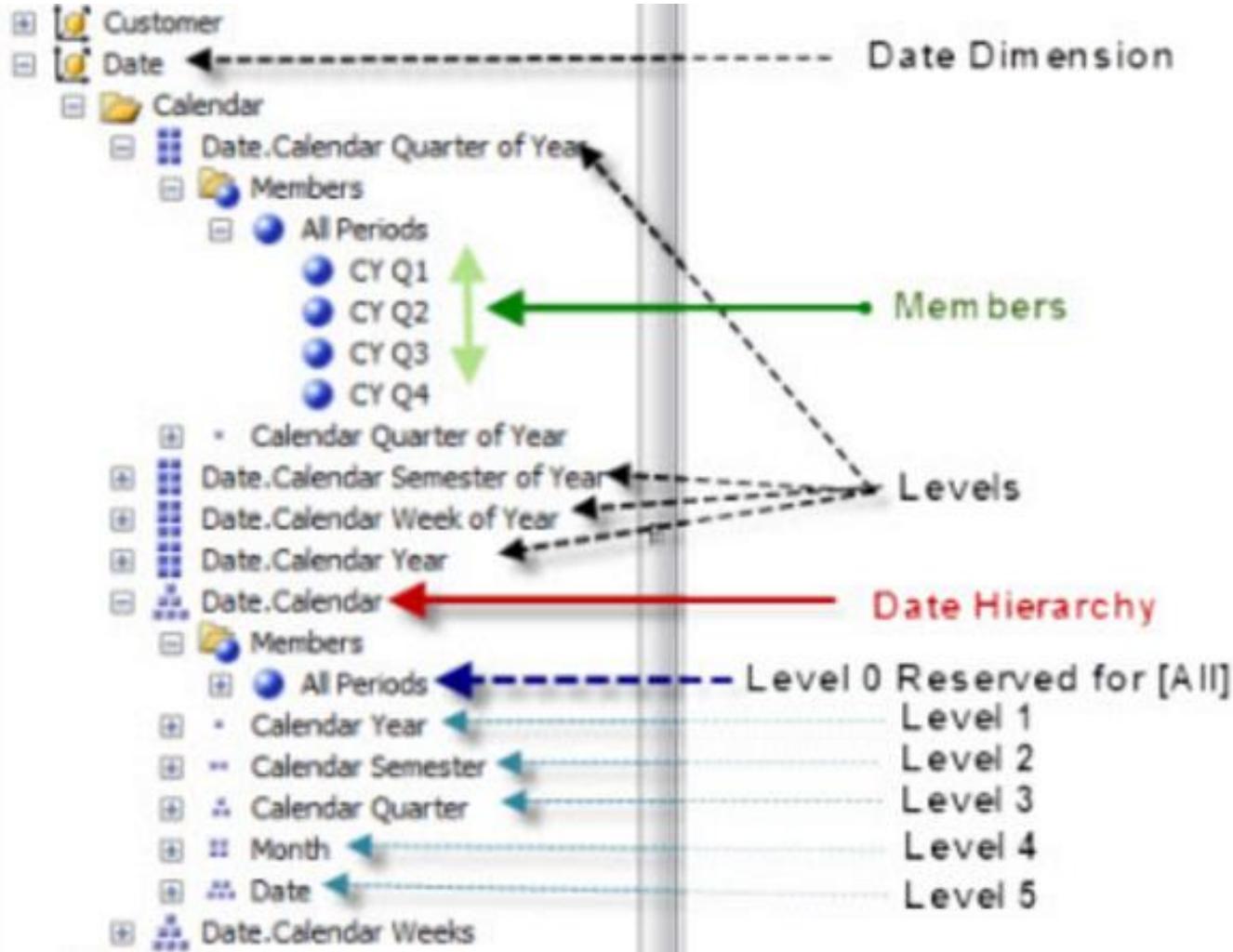
*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Podstawowe funkcje MDX

- Podstawowe grupy funkcji:
  - funkcje tekstowe
  - funkcje dotyczące elementów
  - funkcje numeryczne
- Zastosowanie funkcji
  - podstawowa składnia



„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”



*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Podstawowe funkcje MDX Members

- Zwraca zbiór wszystkich elementów, np. poziomu
- dozwolone jest użycie jako wierszy i/lub jako kolumn

Product.Category.Members



Fundusze  
Europejskie  
Wiedza Edukacja Rozwój



Politechnika Wrocławskiego

Unia Europejska  
Europejski Fundusz Społeczny



*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Podstawowe funkcje MDX

## CurrentMember

- Zwraca bieżący element wymiaru
  - może być etykietą wiersza, lub kolumny, lub filtra
  - jest domyślną funkcją dla wymiaru
  - należy do „Member Group” na liście funkcji

Product.CurrentMember.Name



*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Podstawowe funkcje MDX

## Name

- Zwraca nazwę obiektu jako tekst
- Posiada trzy wersje
  - Name – dla wymiaru lub hierarchii
  - Name – dla poziomu
  - Name – dla elementu
- Należy do „String Group” na liście funkcji

Product.CurrentMember.Name

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Podstawowe funkcje MDX

## Level

- Zwraca poziom dla elementu
- Zwykle następuje po funkcji CurrentMember
- Współpracuje z funkcją Name
- Należy do „Level Group” na liście funkcji

Product.CurrentMember.Level.Name

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Podstawowe funkcje MDX

## Parent

- Zwraca rodzica dla podanego elementu
- Zwykle następuje po funkcji CurrentMember
- Należy do „Member Group” na liście funkcji

Product.CurrentMember.Parent.Name

# Podstawowe funkcje MDX

## Ancestor

- Zwraca przodka elementu
  - na określonym poziomie
  - na określonej „odległości” w hierarchii
- równoważna funkcji Parent dla ...

```
Ancestor(Product.CurrentMember, Category).Name
```

```
Ancestor(Product.CurrentMember, 1).Name
```

„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”

# Podstawowe funkcje MDX Properties

- Zwraca właściwości dla elementu
- wymaga nazwy właściwości w apostrofach
- zwraca błąd jeśli właściwość nie istnieje
- często używana z funkcją Ancestor w celu otrzymania poprawnego poziomu
- zwraca właściwości jako tekst nawet jeśli wartością jest liczba

```
Product.CurrentMember.Properties("Price")
```



Fundusze  
Europejskie  
Wiedza Edukacja Rozwój



Politechnika Wrocławskiego

Unia Europejska  
Europejski Fundusz Społeczny



„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”

# Podstawowe funkcje MDX

## Operatory warunkowe

- Zwraca wartość logiczną TRUE lub FALSE
- operatory: =, >, <, <>, >=, <=
- operatory logiczne: AND, OR, NOT, XOR

```
Ancestor(Product,Category).Name = „Road Bike” OR  
Ancestor(Product,Subcategory).Name = “Mountain Bike”
```

„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”

## Podstawowe funkcje MDX

### Funkcja IIF

- Zwraca jedną z dwóch wartości w zależności od wyrażenia warunkowego
- Zwraca pierwszą wartość, gdy wyrażenie warunkowe jest prawdziwe, jeśli fałszywe zwraca drugą wartość
- IIF jest „prostym” odpowiednikiem klauzuli IF
- Obie wartości muszą być tego samego typu

```
IIF(Ancestor(Product, Category).Name = "Road Bike", "Bring", "Don't Bring")
```



*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Podstawowe funkcje MDX

## IsEmpty

- Zwraca True jeżeli określone wartości są puste (nieokreślone)
- Należy zwrócić uwagę na problem dzielenia przez zero lub inne błędy
- Należy do „Logical Group” na liście funkcji

```
IsEmpty(Measures.CurrentMember)
```

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Podstawowe funkcje MDX

## PrevMember

- Zwraca poprzedni element na tym samym poziomie
- Dozwolone jest użycie z dowolnym wymiarem, ale najczęściej występuje dla hierarchii czasu
- Należy do „Member Group” na liście funkcji

```
[Order Date].[2014].[May].CurrentMember.PrevMember
```

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Podstawowe funkcje MDX

## ParallelPeriod

- Zwraca odpowiadające element kuzyna na równoległym poziomie
- Domyślnie może być użycie bieżącego elementu dla wymiaru czasu
- Należy do „Member Group” na liście funkcji

ParallelPeriod([Order Date].[2014].[Q2].[May])



„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”

## Podstawowe funkcje MDX LastPeriods

- Zwraca zbiór elementów dla poprzednich okresów w hierarchii czasu
- Opcjonalny drugi parametr może wskazać poziom odniesienia

[May].LastPeriods(3)

Avg(LastPeriods(3),[LineTotal])

# Podstawowe funkcje MDX

## PeriodsToDate

- Zwraca zbiór elementów na określonym poziomie od elementu pierwszego do bieżącego
- Drugi parametr może określić pkt. odniesienia
- Kalkulacja wartości narastających w okresie

PeriodsToDate(Year)

Sum(PeriodsToDate([May]),[Line Total])

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Podstawowe funkcje MDX

## YTD

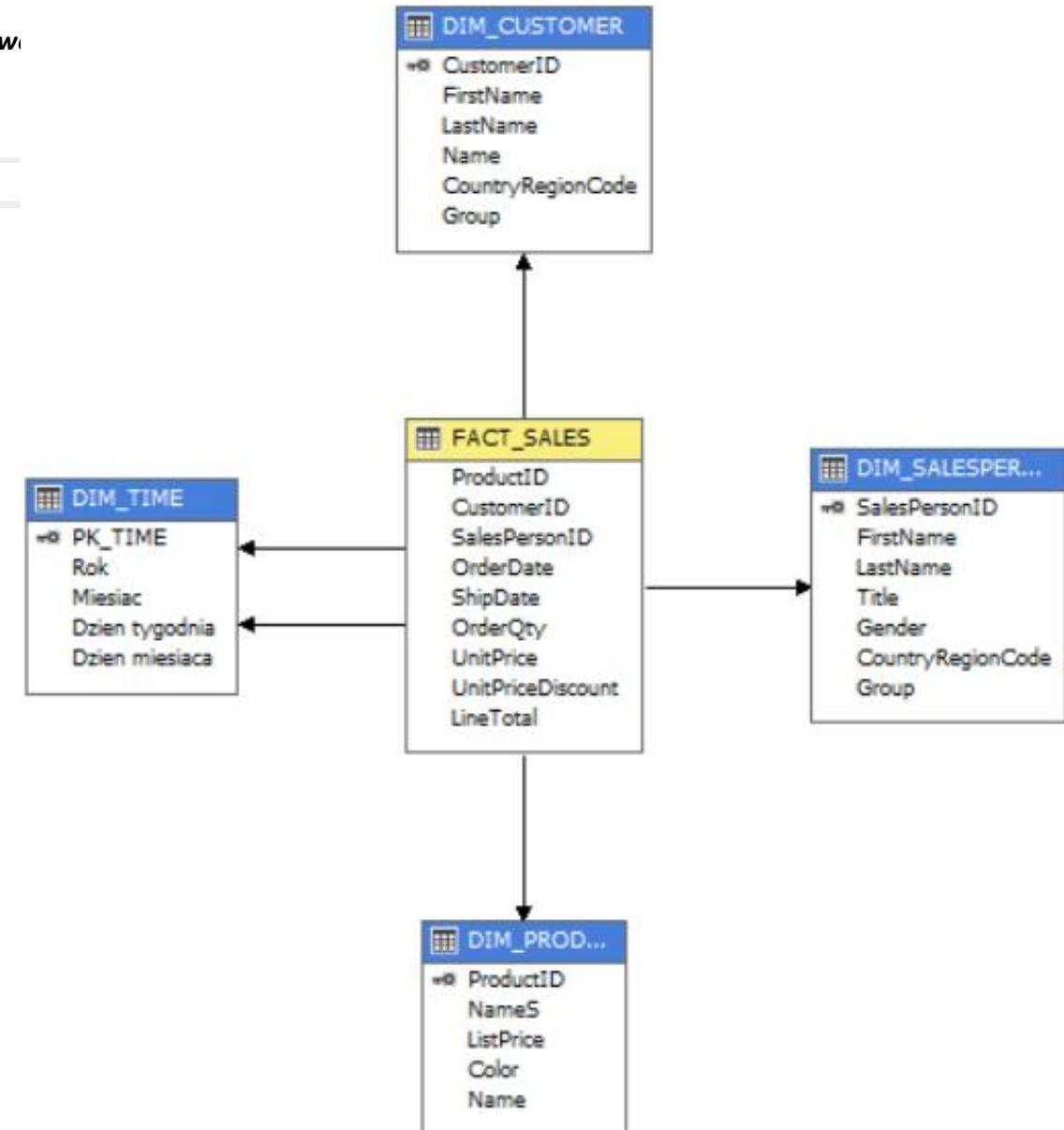
- Zwraca zbiór analogicznie do PeriodsToDate(Year)
- Wymaga deklaracji poziomu Years w wymiarze czasu

Sum(YTD(), [Line Total])



„ZPR PWr – Zintegrowany Program Rozwoju”

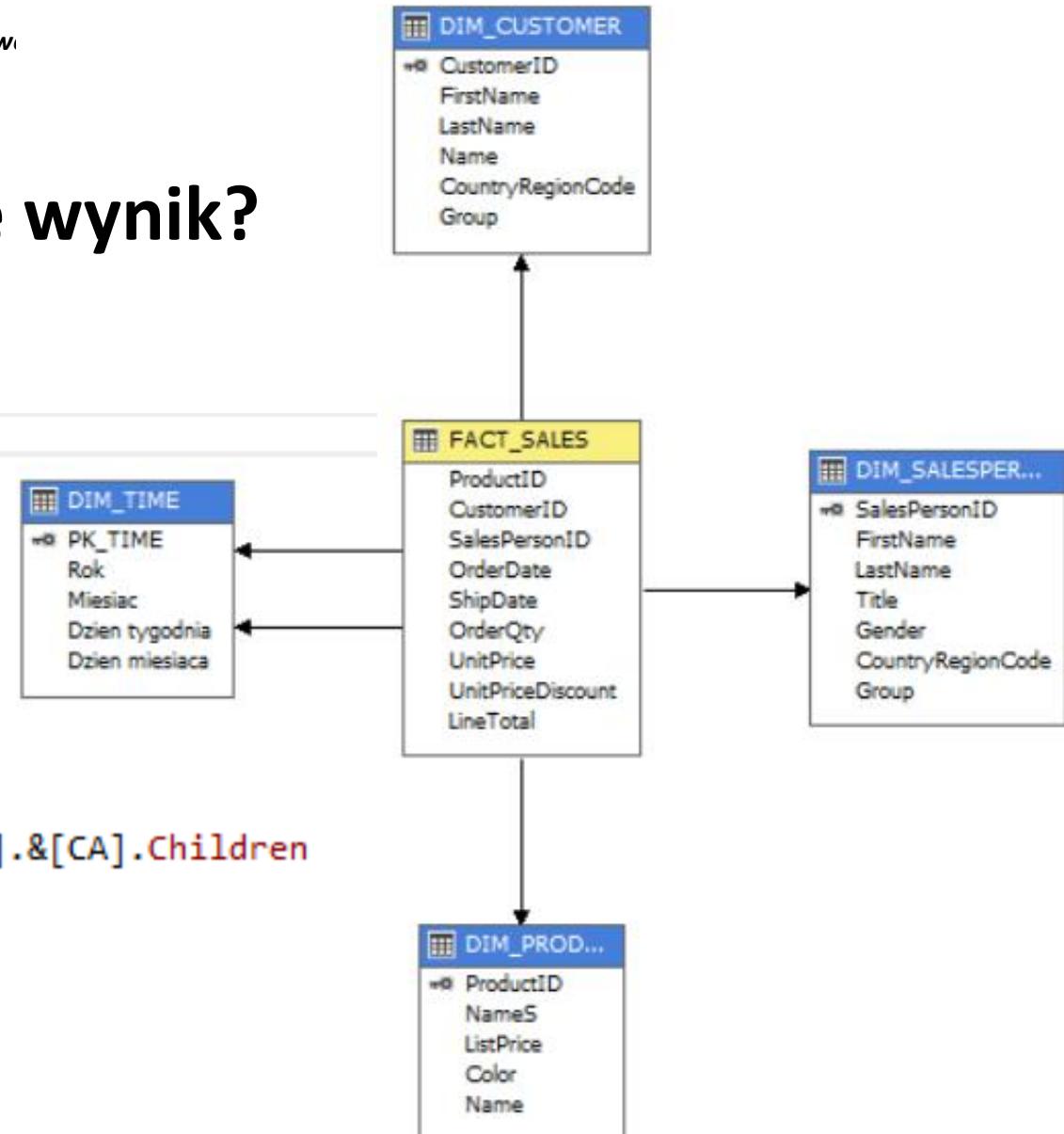
```
SELECT {  
    [DIM TIME].[Rok].&[2012],  
    [DIM TIME].[Rok].&[2013]  
}  
ON COLUMNS,  
{  
    [DIM SALESPERSON].[Country Region Code].&[US],  
    [DIM SALESPERSON].[Country Region Code].&[CA],  
    [DIM SALESPERSON].[Country Region Code].&[FR],  
    [DIM SALESPERSON].[Country Region Code].&[AU],  
    [DIM SALESPERSON].[Country Region Code].&[DE],  
    [DIM SALESPERSON].[Country Region Code].&[GB]  
}  
ON ROWS  
FROM [Adventure Works2014]  
WHERE [Measures].[Sales Person ID Distinct Count]
```





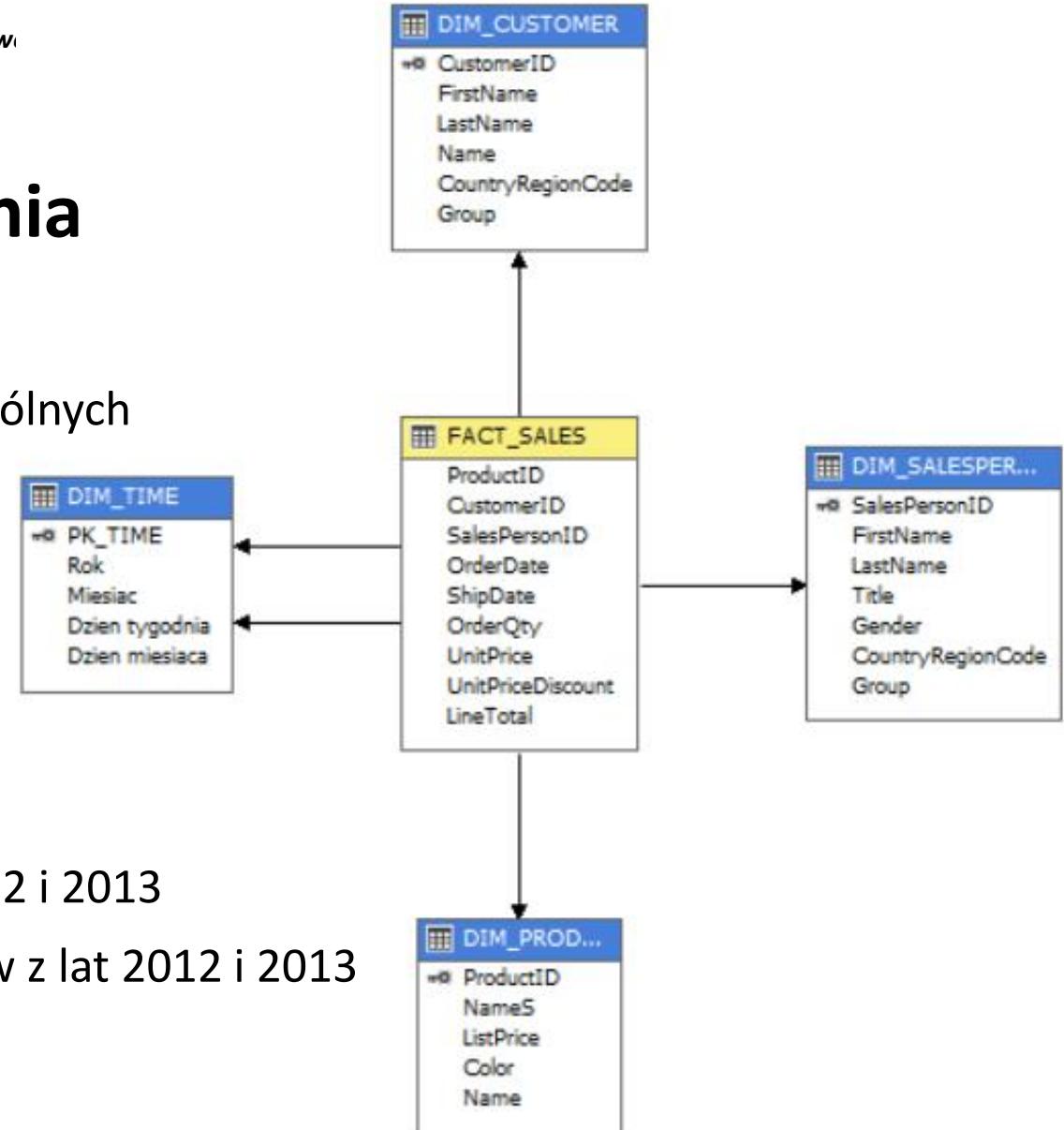
# Jaki będzie wynik?

```
SELECT CROSSJOIN({  
    [DIM TIME].[Rok].&[2012],  
    [DIM TIME].[Rok].&[2013]  
}, {  
    [Measures].[Customer ID Distinct Count],  
    [Measures].[Order Qty]  
})  
ON COLUMNS,  
{  
    [DIM CUSTOMER].[Hierarchy].[Group].&[North America].&[CA].Children  
}  
ON ROWS  
FROM [Adventure Works2014]
```



# Zadania

- Przygotuj zestawienie: liczba transakcji w poszczególnych dniach tygodnia dla różnych kolorów produktów
- Przygotuj zestawienie: liczba zakupionych produktów w poszczególnych miesiącach dla różnych regionów klienta
- Porównaj średnie kroczące dla kwartałów z lat 2012 i 2013
- Porównaj narastające sumy wartości dla kwartałów z lat 2012 i 2013





Fundusze  
Europejskie  
Wiedza Edukacja Rozwój



Politechnika Wrocławska

Unia Europejska  
Europejski Fundusz Społeczny



*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Hurtownie danych

Dziękuję za uwagę

dr inż. Marcin Maleszka



Fundusze  
Europejskie  
Wiedza Edukacja Rozwój



Politechnika Wrocławska

Unia Europejska  
Europejski Fundusz Społeczny



*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Hurtownie danych

**Podstawy MDX**

**dr inż. Marcin Maleszka**

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Tworzenie i operacje na zbiorach

- Funkcje tworzące (zwracające) zbiory
- Funkcje manipulacji na zbiorach
- Funkcje operujące na podzapytaniach
- Operacje na zbiorach elementów wymiaru

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Tworzenie i operacje na zbiorach

- Funkcje tworzące (zwracające) zbiory:
  - Members
  - Children
  - Descendants
- Funkcje manipulacji na zbiorach
- Funkcje operujące na podzapytaniach
- Operacje na zbiorach elementów wymiaru



„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”

## Funkcja Descendants

- Zwraca zbiór elementów na określonym poziomie lub odległości
- Analogiczna do funkcji Ancestor, tylko przenosi „w dół”
- Jeśli został pominięty poziom, to zwraca wszystkich potomków na wszystkich poziomach

Descendants([2016],[Month])

Descendants([2016])

Descendants([2016], , LEAVES)



„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”

# Funkcja Descendants

## Wartości parametru FLAG

Flag	Description
SELF	Returns only descendant members from the specified level or at the specified distance. The function includes the specified member, if the specified level is the level of the specified member.
AFTER	Returns descendant members from all levels subordinate to the specified level or distance.
BEFORE	Returns descendant members from all levels between the specified member and the specified level, or at the specified distance. It includes the specified member, but does not include members from the specified level or distance.
BEFORE_AND_AFTER	Returns descendant members from all levels subordinate to the level of the specified member. It includes the specified member, but does not include members from the specified level or at the specified distance.
SELF_AND_AFTER	Returns descendant members from the specified level or at the specified distance and all levels subordinate to the specified level, or at the specified distance.
SELF_AND_BEFORE	Returns descendant members from the specified level or at the specified distance, and from all levels between the specified member and the specified level, or at the specified distance, including the specified member.
SELF_BEFORE_AFTER	Returns descendant members from all levels subordinate to the level of the specified member, and includes the specified member.
LEAVES	Returns leaf descendant members between the specified member and the specified level, or at the specified distance.

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Funkcje manipulacji na zbiorach

- Funkcje Head i Tail
- Funkcja Union
- Inne funkcje manipulacji na zbiorach
  - Funkcja Intersect
  - Funkcja Except
- Funkcja Hierarchize



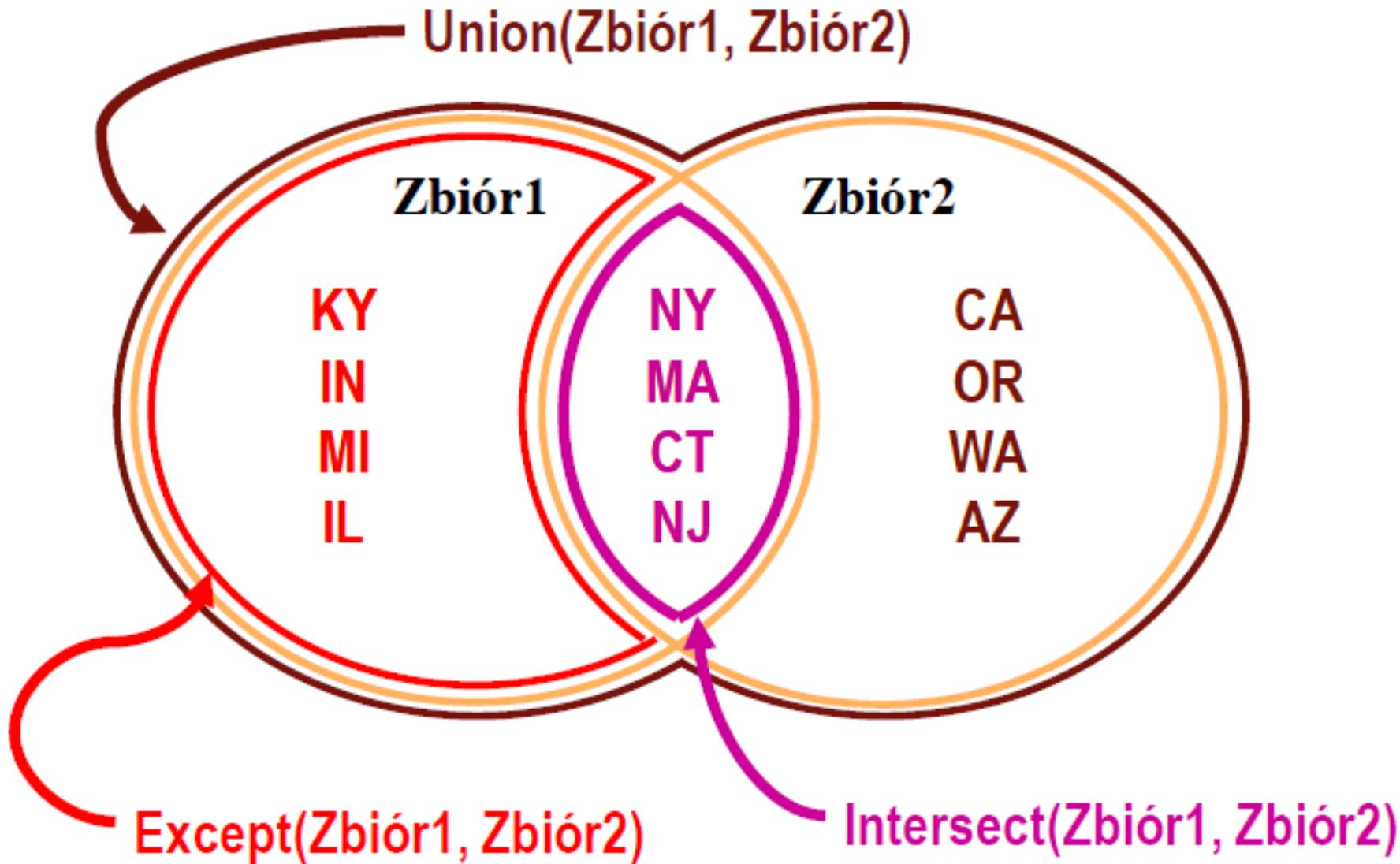
*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

## Funkcja Union

- Zwraca zbiór będący sumą dwóch zbiorów
- Domyślnie pomijane są duplikaty; Opcja ALL pozwana na przywrócenie duplikatów
- Równoważne do zamknięcia zapisu dwóch zbiorów w nawiasach

```
Union([Calendar Year].Members,  
[Calendar Quarter].Members)
```

„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”



*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

## Funkcja Hierarchize

- Zwraca zbiór zorganizowany w kolejności hierarchii
- Rozwijanie hierarchii odbywa się według określonej kolejności zbiorów

```
Hierarchize(Union([Year].Members, [Quarter].Members)))
```



*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Funkcje operujące na podzapytaniach Order

- Zwraca uporządkowany zbiór zgodnie z podzapytaniem
- Możliwe jest użycie numerycznego lub tekstowego wyrażania w podzapytaniu
- Sortowanie zbiorów jest niezależne od innych wymiarów

```
Order(Product.Members,[Line Total],DESC)
```

```
Order(Subcategory.Members,[Line Total],BDESC)
```

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Funkcje operujące na podzapytaniach **TopCount**

- Zwraca określoną liczbę elementów zgodnie z ustaloną kolejnością
- Kombinacja funkcji Order i Head

```
TopCount(Category.Members, 3, [Line Total])
```

# Funkcje operujące na podzapytaniach

## Filter

- Zwraca przefiltrowany zbiór w oparciu o podzapytanie
- Wymaga podania zbioru oraz wyrażenia MDX
- Elementy dla których wyrażenie jest nieprawdziwe są pomijane
- Funkcja jest niezależna od zastosowania innych wymiarów

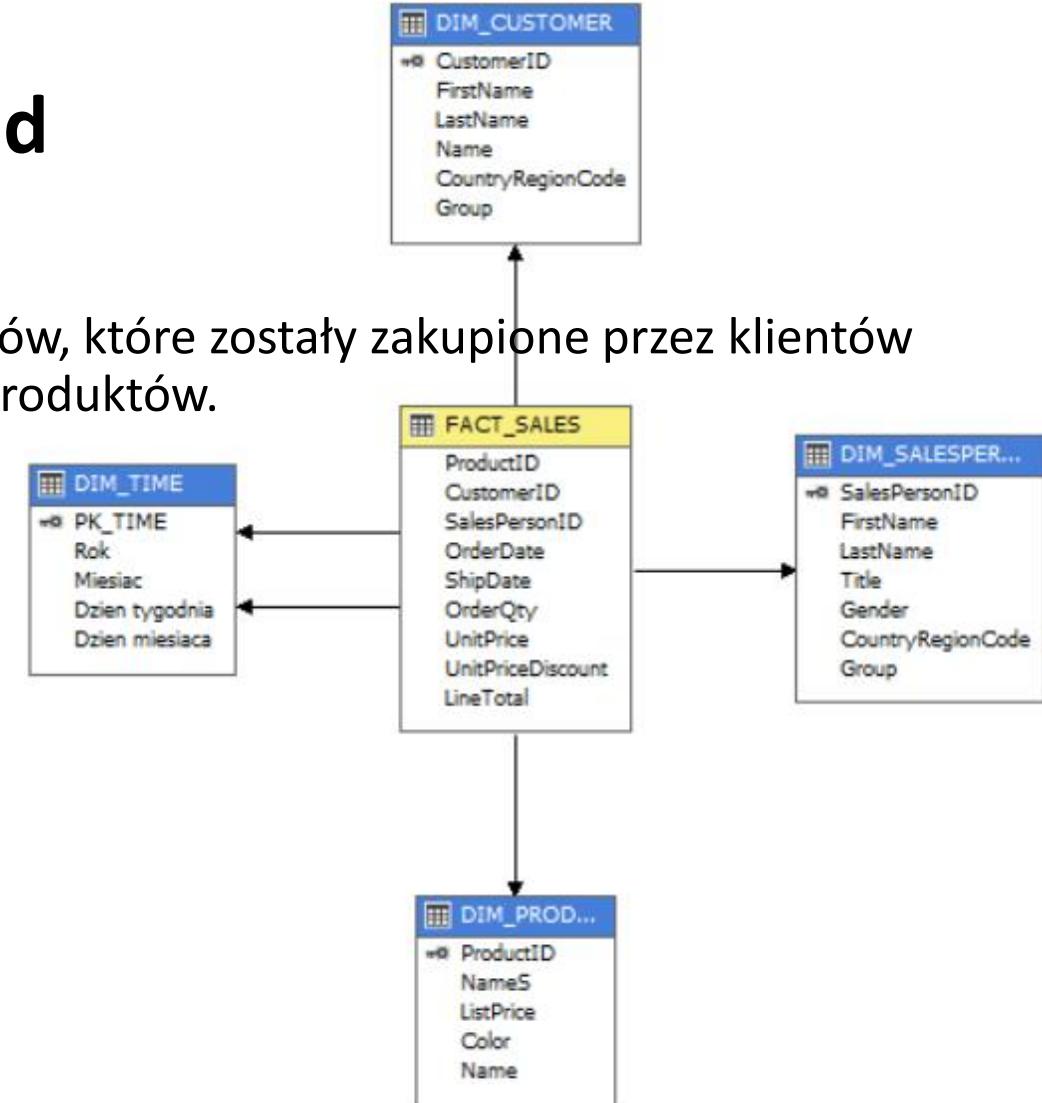
```
Filter([Product Name].Members,[OrderQty] > 1000)
```



# Przykład

- Wyświetl wszystkie kategorie i podkategorie produktów, które zostały zakupione przez klientów z poszczególnych regionów oraz liczbę zakupionych produktów.

```
SELECT CROSSJOIN ({
[DIM CUSTOMER].[Country Region Code].Children
},
{[Measures].[Order Qty]}
)
ON COLUMNS,
[DIM PRODUCT].[Hierarchy].[Category Name],
[DIM PRODUCT].[Sub Category Name].Children
)
ON ROWS FROM [Adventure Works2014];
```





## Przykład

- Wyświetl wszystkie kategorie i podkategorie produktów, które zostały zakupione przez klientów z poszczególnych regionów oraz liczbę zakupionych produktów.
- Wyświetl tylko te podkategorie, w których zakupiono co najmniej 10 różnych produktów

```
SELECT CROSSJOIN ({
[DIM CUSTOMER].[Country Region Code].Children
},
{[Measures].[Order Qty]}
)
ON COLUMNS,
[DIM PRODUCT].[Hierarchy].[Category Name],
ORDER(
    FILTER(
        [DIM PRODUCT].[Sub Category Name].Children,
        [Measures].[Product ID Distinct Count] >= 10
    ),
    [Measures].[Product ID Distinct Count], DESC
)
)
ON ROWS FROM [Adventure Works2014];
```



„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”

```
SELECT CROSSJOIN ({
[DIM CUSTOMER].[Country Region Code].Children
},
{[Measures].[Order Qty]
})
ON COLUMNS,
[DIM PRODUCT].[Hierarchy].[Category Name],
ORDER(
    FILTER(
        [DIM PRODUCT].[Sub Category Name].Children,
        [Measures].[Product ID Distinct Count] >= 10
    ),
    [Measures].[Product ID Distinct Count], DESC
)
)
ON ROWS FROM [Adventure Works2014];
```

## Przykład



```
SELECT CROSSJOIN ({
[DIM CUSTOMER].[Country Region Code].Children
},
{[Measures].[Order Qty]
})
ON COLUMNS,
[DIM PRODUCT].[Hierarchy1].[Category Name]
ORDER(
  FILTER(
    [DIM PROD].[Measure]
  )
),
[Measure]
)
)
ON ROWS FROM [A]
```

## Przykład

		AU	CA	DE	FR	GB	US
		Order Qty					
Accessories	Tires and Tubes	3307	2861	1597	1817	2026	6398
Bikes	Road Bikes	2605	8198	1083	2659	2627	30024
	Mountain Bikes	1430	4596	736	1464	1926	18169
Bikes	Touring Bikes	1993	1620	1691	1615	1387	6445
Components	Mountain Frames	103	2207	170	789	732	7620
Components	Road Frames	1	2653	84	752	509	7754
Components	Touring Frames	282	489	473	408	323	1750
Components	Wheels	(null)	1035	10	251	286	3691

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# **Operacje na zbiorach elementów wymiaru NON EMPTY**

- Ukrywa puste elementy (wiersze lub kolumny)
- Umieszczana bezpośrednio przed określeniem zbioru

NON EMPTY Product.Members ON ROWS

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Operacje na zbiorach elementów wymiaru **CROSSJOIN**

- Zwraca pojedynczy zbiór będący kombinacją dwóch różnych zbiorów z różnych wymiarów
- Każdy element drugiego zbioru występuje dla każdego elementu z pierwszego zbioru

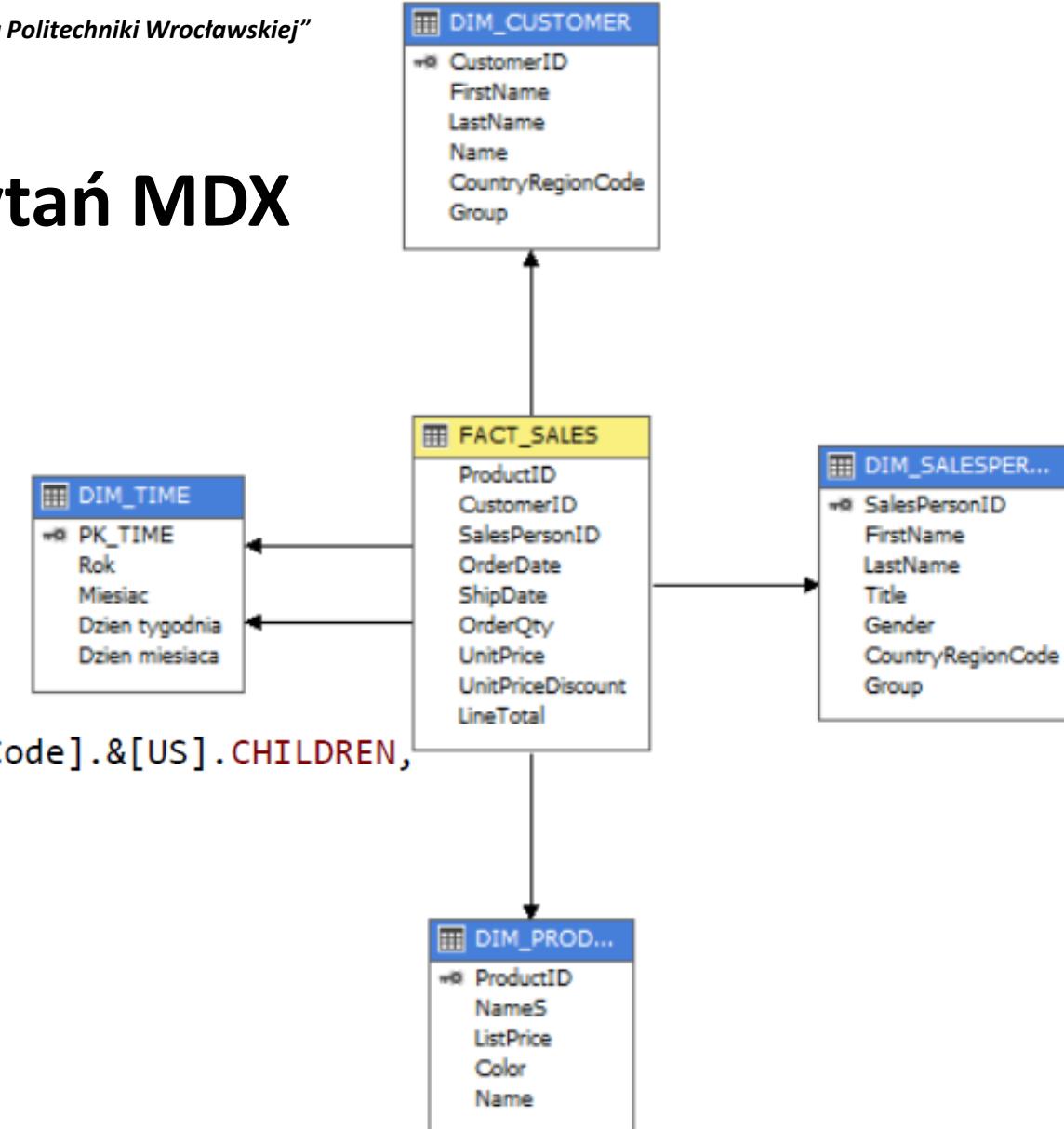
```
Crossjoin([Calendar Year].Members, [Category].Members)
```



„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”

# Przykłady zapytań MDX

```
select
{
    [Measures].[Order Qty],
    [Measures].[Sales Person ID Distinct Count]
} on columns,
order
(
    filter
    (
        [DIM CUSTOMER].[Hierarchy].[Country Region Code].&[US].CHILDREN,
        [Measures].[Order Qty] > 4
    ),
    [Measures].[Order Qty],
    DESC
) on rows
from [Adventure Works2014]
where [Order Date].[Year].[2013];
```



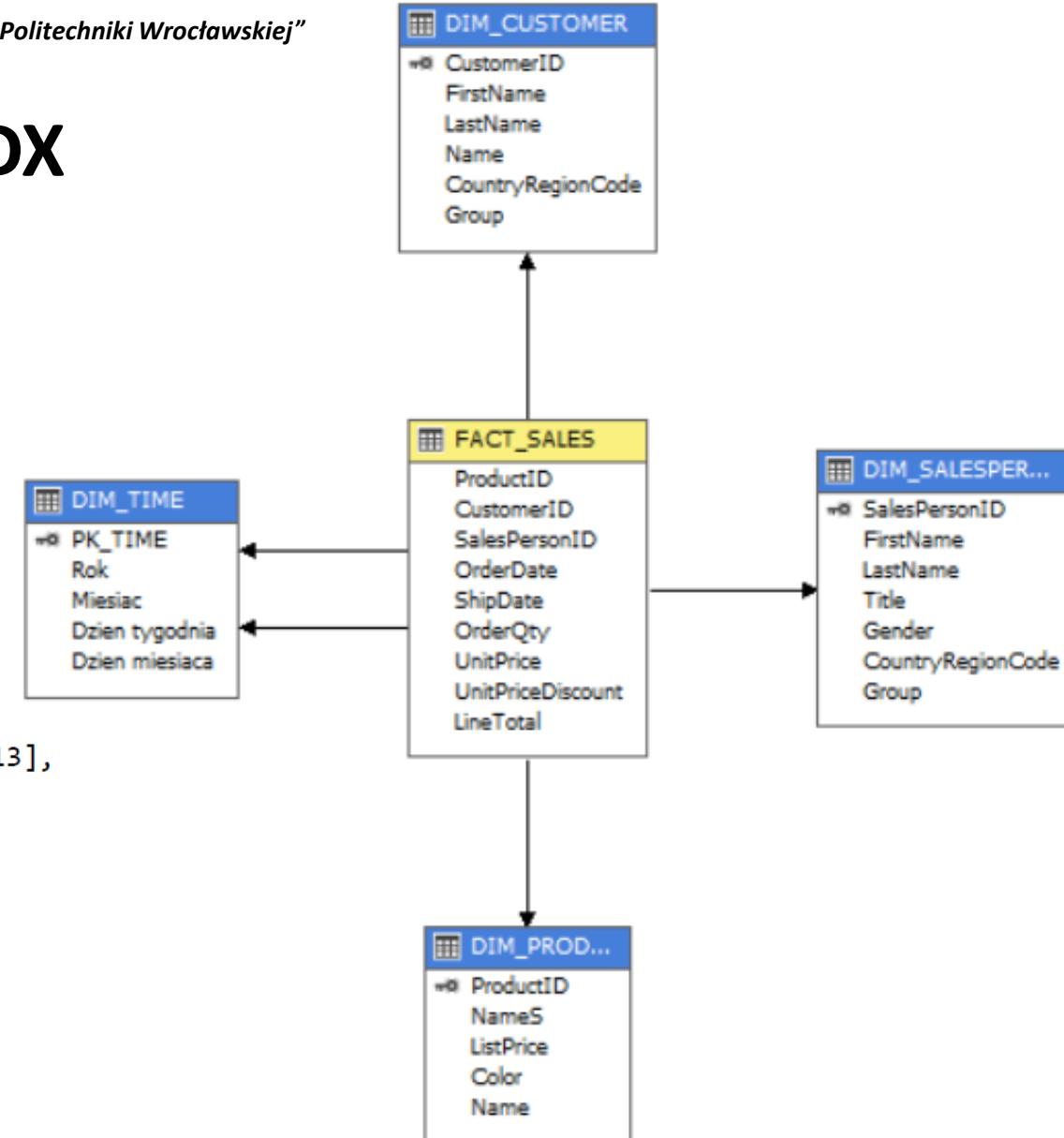


„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”

# Przykłady zapytań MDX

```
with member [Measures].[NazwaMiesiąca]
    as '[ORDER DATE].[Month].CurrentMember.Name'

select
{
    [Measures].[NazwaMiesiąca],
    [Measures].[Order Qty]
} on columns,
head
(
    order
    (
        descendants
        (
            [Order Date].[Hierarchy].[Year].[2013],
            [Order Date].[Hierarchy].[Month],
            AFTER
        ),
        [Measures].[Order Qty],
        BDESC
    )
)
on rows
from [Adventure Works2014];
```





„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”

## Kolejność wartości wymiaru

```
WITH MEMBER [Measures].[Kolejnosc]
AS CASE [DIM TIME].[Miesiąc].CurrentMember
    WHEN [DIM TIME].[Miesiąc].&[Styczeń] THEN 1
    WHEN [DIM TIME].[Miesiąc].&[Luty] THEN 2
    WHEN [DIM TIME].[Miesiąc].&[Marzec] THEN 3
    WHEN [DIM TIME].[Miesiąc].&[Kwiecień] THEN 4
    WHEN [DIM TIME].[Miesiąc].&[Maj] THEN 5
    WHEN [DIM TIME].[Miesiąc].&[Czerwiec] THEN 6
    WHEN [DIM TIME].[Miesiąc].&[Lipiec] THEN 7
    WHEN [DIM TIME].[Miesiąc].&[Sierpień] THEN 8
    WHEN [DIM TIME].[Miesiąc].&[Wrzesień] THEN 9
    WHEN [DIM TIME].[Miesiąc].&[Październik] THEN 10
    WHEN [DIM TIME].[Miesiąc].&[Listopad] THEN 11
    WHEN [DIM TIME].[Miesiąc].&[Grudzień] THEN 12
END
```

	2012	2013
	SredniaKroczaca	SredniaKroczaca
Styczeń	3970627.28	2087872.46
Luty	1674064.44	2814526.2
Marzec	2508592.65	2991538.29
Kwiecień	2854453.32	3896273.89
Maj	2295057.27	3260659.16
Czerwiec	4099354.36	5081069.13
Lipiec	2995328.46	4146293.85
Sierpień	2396345.37	3480642.77
Wrzesień	2773437.93	3548184.15
Październik	2658209.43	3740859.31
Listopad	2770803.16	3979461.13
Grudzień	3464379.59	4578277.88



Fundusze  
Europejskie  
Wiedza Edukacja Rozwój



Politechnika Wrocławska

Unia Europejska  
Europejski Fundusz Społeczny



*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Hurtownie danych

Dziękuję za uwagę

dr inż. Marcin Maleszka



Fundusze  
Europejskie  
Wiedza Edukacja Rozwój



Politechnika Wrocławskiego

Unia Europejska  
Europejski Fundusz Społeczny



*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Hurtownie danych

**Wielowymiarowy model danych - warstwa fizyczna**

**dr inż. Marcin Maleszka**

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Modelowanie hurtowni danych przypomnienie

- Model biznesowy
  - Efekt analizy strategicznej
  - Identyfikacja miar i wymiarów dla poszczególnych procesów biznesowych
- Model logiczny (wymiarowy)
  - Model abstrakcyjny, konceptualny
  - Encje i atrybuty (reprezentowane w modelu relacyjnym jako tabele i powiązania między nimi)

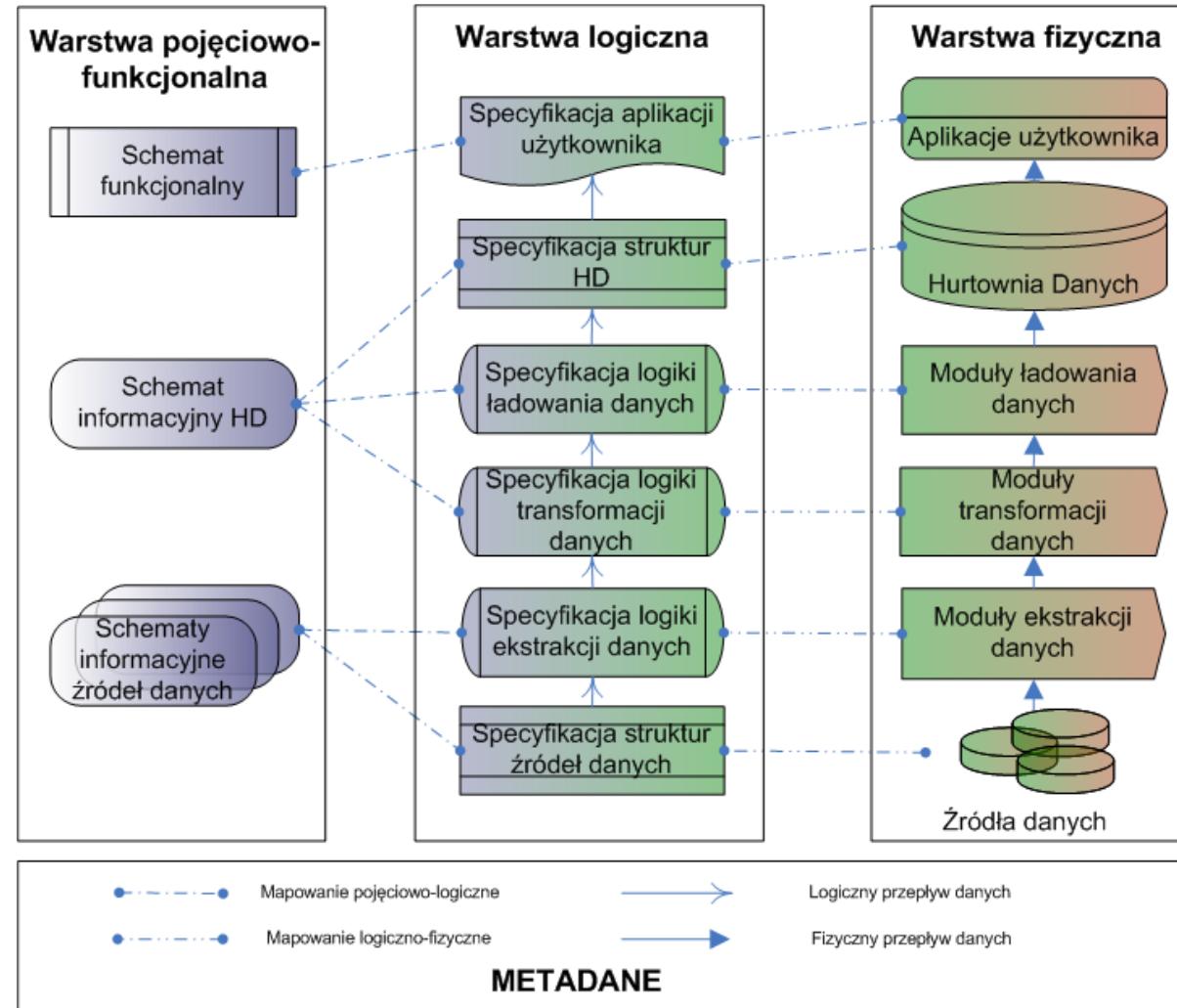
*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Modelowanie hurtowni danych przypomnienie

- Model fizyczny
  - Wybór sposobu składowania danych
  - Formaty danych
  - Strategie partycjonowania
  - Wybór indeksów
  - Wybór materializowanych perspektyw



„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”



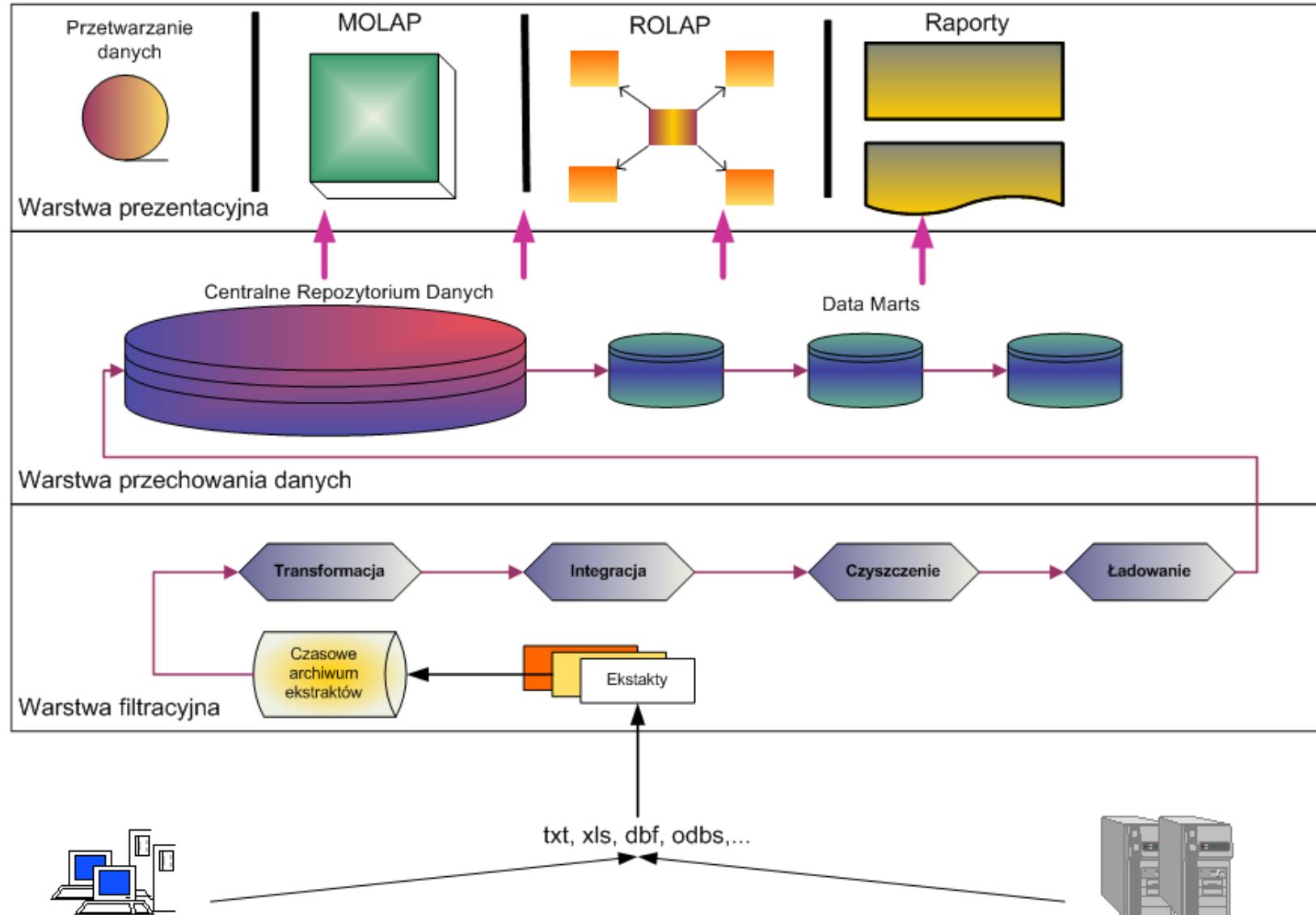
*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Fizyczny projekt i rozwój bazy danych

- projektowanie bazy danych
- identyfikacja kluczy
- przygotowanie strategii agregacji danych
- tworzenie strategii indeksowania
- przygotowanie strategii podziału (partycjonowania)
- planowanie pojemności (strategia gromadzenia)
- tworzenie obiektów bazy danych

# Warstwa fizyczna

- aplikacje użytkownika
- przechowywanie danych
- filtracja danych
  - transformacja
  - ładowanie
- źródła danych
  - ekstrakcja



*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Ekstrakcja ze źródeł danych

- zintegrowany proces pozyskania danych z różnych źródeł
- dostępność danych
- wydajność procesu ekstrakcji (bez zakłócania funkcjonowania systemów źródłowych)
- weryfikacja poprawności danych
- automatyzacja procesu

# Filtracja

- Odczyt danych z systemów źródłowych
  - identyfikacja nadania danych i weryfikacja ich źródła
  - weryfikacja kompletności danych
  - wczytanie do struktur pomocniczych
  - przechowywanie danych źródłowych do momentu zakończenia procesu ładowania
  - dokumentacja procedury (logi operacji, listy kontrolne)
- Transformacja
  - konwersja danych ze struktur tymczasowych (ang. TSA – Temporary Storage Area)
    - konwersja typów i wartości
    - kategoryzacja wartości
    - sprawdzenie poprawności



*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Filtracja

- Integracja
  - relacje pomiędzy danymi z różnych źródeł
  - agregacja danych
- Czyszczenie
  - sprawdzenie poprawności merytorycznej danych z różnych źródeł
  - reguły kontrolne, np. sumy krzyżowe, zgodność sum, kompletność relacji
  - reguły naprawcze dla wykrytych niespójności
- Ładowanie danych
  - przepisanie danych z TSA do stałych struktur HD – CRD (ang. Central Repository of Data)
  - działania sterowane metadanymi procesu

# Przechowywanie danych - Centralne repozytorium danych

- Implementacja relacyjna
  - model gwiazdy, płatka śniegu, konstelacji faktów
  - perspektywy zmaterializowane
  - partycjonowanie danych
- Implementacja wielowymiarowa
  - wielowymiarowe kostki
  - selekcja i wycinanie (ang. slice & dice)
  - zwijanie (ang. roll-up, drill-up)
  - drążenie (ang. drill-down, drill-through)
  - obracanie (ang.pivoting)



*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Przechowywanie danych – Data Marts

- Hurtownie tematyczne – kopia wybranych danych z CRD
- Mniejsze dane -> lepsza wydajność, możliwość rozproszenia, częściej model wielowymiarowy
- Zakres wyodrębnionych danych definiowany przez potrzeby użytkowników:
  - sprawozdawcze
  - raportowe
  - monitorowanie poziomu realizacji planu
  - analityczne (analiza statystyczna lub eksploracja danych)
  - CRM
  - itp..

# Widoki/Perspektywy zmaterializowane

- Problem:
  - Duża tabela faktów + mniejsze wymiary
  - Optymalizacja odczytu, a złączenia kosztowne
- Cel:
  - Przyspieszenie odczytu
- Rozwiązanie
  - Unikanie wielokrotnego wykonywania tych samych operacji
  - Widok zmaterializowany



*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Widok zmaterializowany

```
CREATE MATERIALIZED VIEW Nazwa_widoku AS  
SELECT ... FROM ... WHERE Warunek
```

Zapytania do widoku:

```
SELECT ... FROM Nazwa_widoku WHERE P1;
```

```
SELECT ... FROM Nazwa_widoku WHERE P2;
```

...

```
SELECT ... FROM Nazwa_widoku WHERE PN;
```



*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Widoki zmaterializowane

- Korzyści:
  - Dla zwykłego widoku – niejawne wykonanie dla każdego zapytania
  - Wyniki widoku zmaterializowanego przechowywane są fizycznie w systemie
  - Można użyć tabeli tymczasowej
- Ograniczenia:
  - Niebezpieczeństwo niespójności danych (opóźnienie dla automatycznego odświeżania)
  - Wyniki oparte na nieaktualnych danych
  - Czy w HD te aspekty są ważne?

## Optymalizacja z wykorzystaniem widoków zmaterializowanych

- Perspektywa zmaterializowana przechowuje wyniki czasochłonnych zapytań analitycznych
- Wydajne, jeśli odpowiedź na zapytanie identyczne lub podobne do zapytania definiującego perspektywę
- Dane w tabelach źródłowych perspektywy nie ulegają modyfikacji
- Sposoby:
  - przepisywanie zapytań
  - indeksowanie

# Indeksowanie

- Indeks połączeniowy (ang. join index)
  - łączy z sobą rekordy z różnych tabel posiadające tę samą wartość atrybutu połączeniowego
  - posiada strukturę B-drzewa zbudowanego na atrybucie połączeniowym tabeli
  - liście indeksu zawierają wspólne wartości atrybutu połączeniowego tabel wraz z listami adresów rekordów w każdej z łączonych tabel
- Indeks bitmapowy (ang. bitmap index)
  - wykorzystanie pojedynczych bitów do zapamiętania informacji o tym, że dana wartość atrybutu występuje w określonym rekordzie tabeli
  - mapa bitowa reprezentująca każdą unikalną wartość atrybutu
  - każdy bit mapy odpowiada jednemu rekordowi w tabeli
  - zbiór map bitowych dla danego atrybutu
  - B-drzewo z mapami bitowymi w liściach

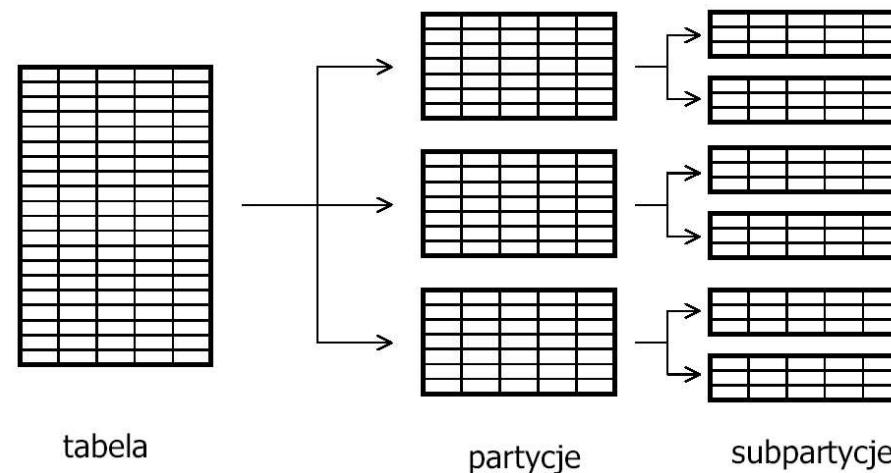
*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Indeksowanie

- Bitmapowy indeks połączeniowy (ang. bitmap join index)
  - połączenie indeksu połączeniowego i bitmapowego
  - tworzenie tylu map bitowych ile jest wartości unikalnych atrybutu
  - każda mapa opisuje rekordy z tabeli
  - struktura B-drzewa, w którego liściach znajdują się mapy bitowe opisujące łączone rekordy

# Partycjonowanie

- Fizyczny podział danych na niewielkie, łatwe w zarządzaniu podzbiory, nazywane partycjami
- Każda partycja stanowi odrębny segment w bazie danych
- Partycje mogą być opcjonalnie dzielone na subpartycje
- Partycjonowanie umożliwia równoległą realizację poleceń DML



# Metody partycjonowania

## Partycjonowanie zakresowe

- podział według przynależności wartości kolumny-klucza do predefiniowanych przedziałów

## Partycjonowanie haszowe

- podział według wartości funkcji haszowej (modulo) wyliczanej dla kolumny-klucza

## Partycjonowanie wg listy

- podział według przynależności wartości kolumny-klucza do predefiniowanych list wartości

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Metody partycjonowania

## Partycjonowanie dwupoziomowe zakresowo-haszowe

- rozdział rekordów na partycje wg zakresów, a następnie na subpartycje wg wartości funkcji haszowej

## Partycjonowanie dwupoziomowe zakresowo-listowe

- rozdział rekordów na partycje wg zakresów, a następnie na subpartycje wg przynależności do list wartości

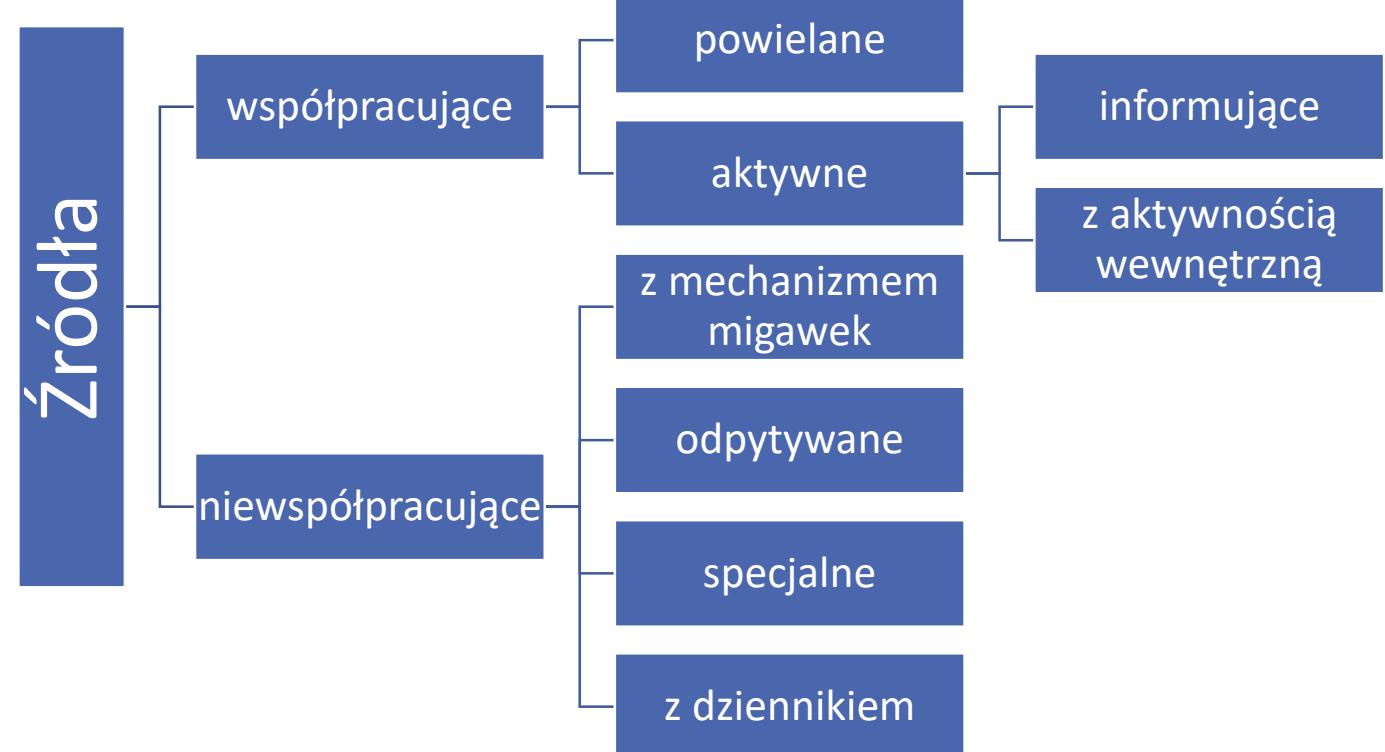
*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

## Zalety partycjonowania

- zrównoleglenie operacji dostępu do dysku, zapytań SQL do różnych partycji
  - równoważenie obciążenie dysków
  - przyspieszenie działania poprzez zapytania do konkretnej partycji
  - bezpieczeństwo danych w razie awarii sprzętowych
  - po awarii niezbędne odtworzenie partycji, a nie całej tabeli
- 
- Wady?

# Odświeżanie danych

- Zapewnienie zgodności danych w hurtowni z danymi źródłowymi
- Wykrywanie zmian w danych źródłowych:
  - monitor zmian
- Klasyfikacja źródeł danych:



# Propagacja aktualizacji

- Nanoszenie zmian na wszelkie potworzone materializacje np. agregacje, hurtownie tematyczne, czy kostki OLAP
- Strategie:
  - Aktualizacja opóźniona (na żądanie, przy pierwszym użyciu po zmianie danych w hurtowni):
    - dłużej trwa pierwsze zapytanie,
    - nie musimy odświeżać tych perspektyw, których nie użyjemy.
  - Aktualizacja natychmiastowa (podczas odświeżania hurtowni):
    - dłużej trwa wsadowe przetwarzanie procesu aktualizacji,
    - przerzucamy kosztowne procesy na godziny nocne,
    - część aktualizacji może okazać się zbędna.



Fundusze  
Europejskie  
Wiedza Edukacja Rozwój



Politechnika Wrocławskiego

Unia Europejska  
Europejski Fundusz Społeczny



*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Prezentacja

- Mechanizmy dostępu i uprawnień użytkowników
- Wybór najlepszego rodzaju OLAP:
  - ROLAP
  - MOLAP
  - HOLAP
- Interfejs
- Raporty i zestawienia statyczne i dynamiczne



Fundusze  
Europejskie  
Wiedza Edukacja Rozwój



Politechnika Wrocławska

Unia Europejska  
Europejski Fundusz Społeczny



*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Hurtownie danych

Dziękuję za uwagę

dr inż. Marcin Maleszka



Fundusze  
Europejskie  
Wiedza Edukacja Rozwój



Politechnika Wrocławska

Unia Europejska  
Europejski Fundusz Społeczny



*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Hurtownie danych

**Podstawy raportowania**

**dr inż. Marcin Maleszka**



*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Raporty

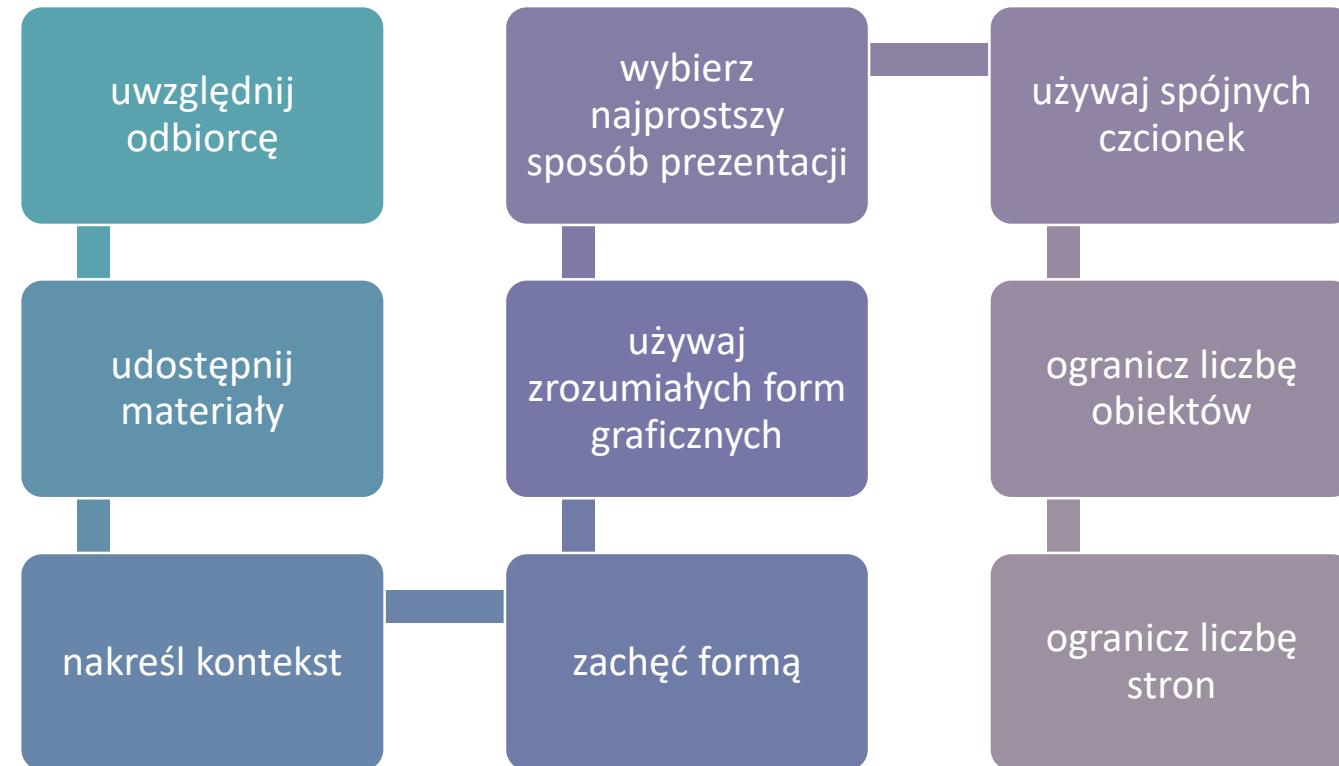
- Definicja raportu:
  - dokument dostarczający niezbędnych informacji decydentom w celu wsparcia ich pracy
- Cel:
  - wykorzystanie informacji w celu redukcji kosztów i zwiększenia zysku (!)
  - pomoc dla managerów w szybkim dostępie do informacji
  - proste zarządzanie
  - interpretacja uzyskanych wyników
  - możliwość wizualizacji danych
- Forma:
  - tekst, tabele, wykresy

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Metodologia



# Wskazówki dla projektantów raportów



*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Typy złożonych raportów

- Zarządzanie pomiarami (ang. metric management)
- Panele nawigacyjne (ang. dashboards)
- Karty wyników (ang. balanced scorecards)
- Analizy Ad-Hoc
- Raporty interaktywne
- Raporty oparte na eksploracji danych i zaawansowanej statystyce



*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Typy złożonych raportów

- Zarządzanie pomiarami
  - wskaźniki zorientowane na wynik
  - gwarantowany poziom usług (SLA) lub kluczowe wskaźniki wydajności (KPI)
  - śledzenie osiągania określonych celów w ustalonym czasie
- Panele nawigacyjne
  - prezentacja wielu wskaźników na jednym ekranie
  - „raporty w puszcze” – statyczne raporty o ustalonej strukturze
  - możliwość dostosowania zawartości do potrzeb
  - możliwość wyróżniania poziomu wskaźników

# Typy złożonych raportów

- Karty wyników
  - metoda opisana przez Kaplana i Nortona
  - prezentacja zintegrowanego widoku sukcesu przedsiębiorstwa
  - głównie wyniki finansowe
  - uwzględnienie innych czynników, np. klienta, procesów biznesowych, perspektyw rozwoju
- Analizy Ad-Hoc
  - zazwyczaj jednorazowe raporty
  - symulacje określonych warunków (co-jeśli)
  - mogą być w formie krótkiego raportu dla kierownictwa

# Typy złożonych raportów

- Raporty interaktywne
  - najczęściej na bazie technologii OLAP
  - pozwalają na interakcję w czasie prezentacji
  - możliwa modyfikacja wymiarów
  - możliwa edycja poziomu ziarnistości (drill-down, roll-up)
- Raporty oparte na eksploracji danych i zaawansowanej statystyce
  - wykorzystanie technik uczenia maszynowego lub sieci neuronowych do odkrywania wzorców w danych
  - klasyfikacja, segmentacja
  - grupowanie
  - predykcja

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

## Elementy składowe

- **Tytuł** – długi opisowy i krótki
- **Identyfikator** – ułatwienie przy odwołaniach
- **Wygląd** – kolory zgodne ze schematem kolorów organizacji, właściwa czcionka, układ, itp.
- **Źródło** – na jakiej podstawie został przygotowany
- **Data** – determinuje stan danych, na podstawie których został przygotowany
- **Właściciel raportu** – ustalenie osoby odpowiedzialnej

## Elementy składowe

- **Opis raportu** – abstrakt i krótka informacja dla kogo raport będzie przydatny, w jakim celu został przygotowany
- **Definicje** – wyjaśnienie używanych pojęć
- **Treść** – tabele, zestawienia, wykresy, diagramy, itp.
- **Notka prawna** – klauzule informacyjne, poufności, itp.
- **Status** – projekt, raport tymczasowy, przyjęty, poprawiony, itp.
- **Dane kontaktowe** – imię i nazwisko oraz kontakt do autora lub osoby, która może udzielić szerszych informacji

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Projektowanie hurtowni a końcowe raporty

- Zgodność celu budowy hurtowni danych z celami biznesowymi przedsiębiorstwa
- Dostępność wszystkich niezbędnych informacji (atrybutów)
- Definicje miar i wskaźników KPI



*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Wymagania

- Wymagania dotyczące końcowych raportów definiują zakres danych i ich analiz:
  - dane rozproszone, często również dane zewnętrzne pochodzące z różnych systemów
  - chcemy badać wpływ tych czynników na wyniki przedsiębiorstwa
  - identyfikacja i ocena jakości danych źródłowych
  - wpływ architektury systemu i raportowania
  - analiza metod i narzędzi wykorzystywanych obecnie
  - potencjalne zestawienia, które można będzie realizować na podstawie hurtowni

**„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”**

# Trendy

- skalowalność
  - wielość i różnorodność źródeł danych
  - hardware, rozwój architektury
- integracja
  - źródeł danych
  - systemów OLTP
  - platformowa i aplikacyjna
  - integracja z sieciami społecznościowymi (dane ustrukturyzowane i nieustrukturyzowane)
- integracja w czasie rzeczywistym i podejmowanie decyzji
- wizualizacja



*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

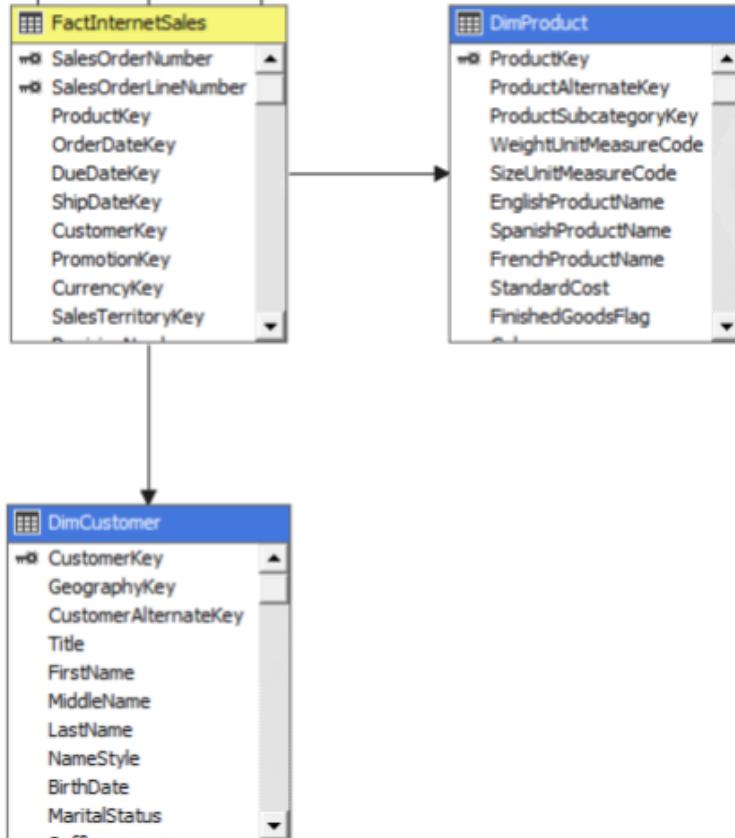
# Elementy eksploracji danych

- Algorytmy klasyfikacyjne
  - przewidywanie wartości dyskretnych na podstawie innych atrybutów w zbiorze danych
- Algorytmy regresyjne
  - predykcja liczbowych wartości ciągłych, np. profit, strata na podstawie innych atrybutów w zbiorze danych
- Algorytmy segmentacji/grupowania
  - podział danych na grupy elementów o podobnych właściwościach
- Algorytmy asocjacyjne (np. reguły asocjacyjne)
  - znajdują korelację pomiędzy atrybutami opisującymi dane
- Analiza sekwencji/trendu
  - znajdowanie powtarzających się sekwencji lub epizodów w danych

DimDate	
▪ DateKey	
▪ FullDateAlternateKey	
▪ DayNumberOfWeek	
▪ EnglishDayNameOfWeek	
▪ SpanishDayNameOfWeek	
▪ FrenchDayNameOfWeek	
▪ DayNumberOfMonth	
▪ DayNumberOfYear	
▪ WeekNumberOfYear	
▪ EnglishMonthName	



## Przykład - kostka

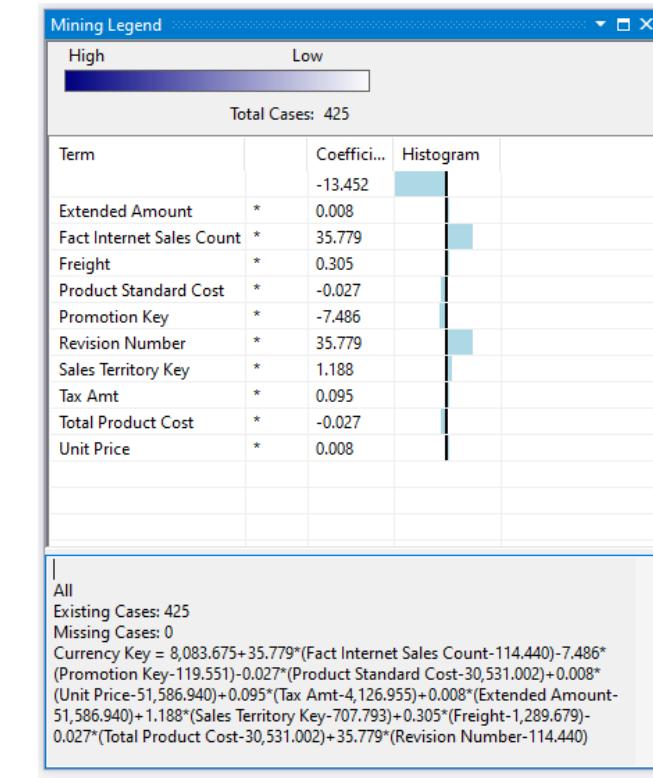
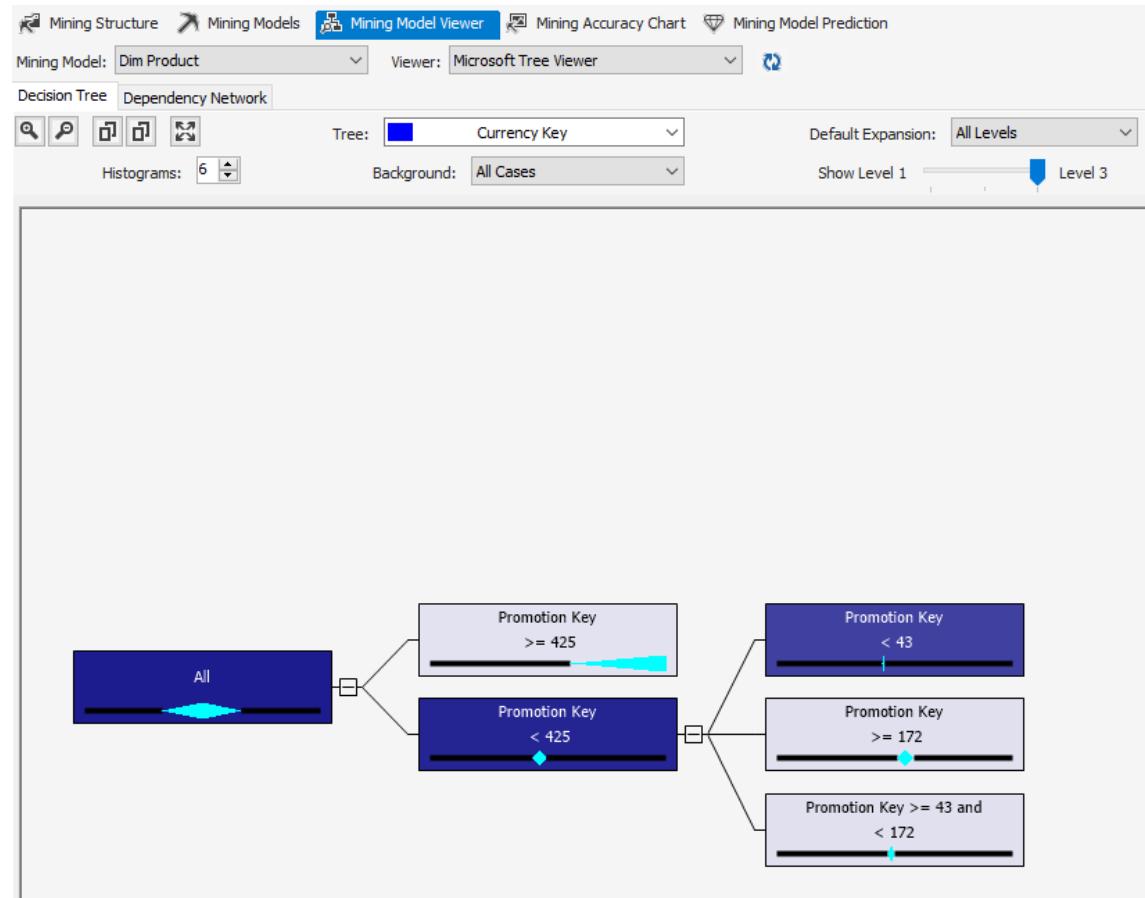


- Wymiary:
  - Produkt
  - Klient
  - Data
- Fakt:
  - transakcja internetowa
- Miary:
  - Liczba zamówionych produktów
  - Kwota transakcji
  - Fracht
  - Liczba transakcji
  - ....



„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”

# Przykład – drzewa decyzyjne





„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”

## Przykład – reguły asocjacyjne

- Na podstawie dostępnych miar przewidujemy wartość miar OrderQty oraz SalesAmount

**Data Mining Wizard**

**Specify Mining Model Column Usage**  
Specify the usage of mining model columns and optionally add nested tables.

**Mining model structure:**

	Tables/Columns	Input	Predict...
<input checked="" type="checkbox"/>	Fact Internet Sales Count	<input checked="" type="checkbox"/>	<input type="checkbox"/>
<input checked="" type="checkbox"/>	Freight	<input checked="" type="checkbox"/>	<input type="checkbox"/>
<input checked="" type="checkbox"/>	Order Quantity	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
<input checked="" type="checkbox"/>	Product Standard Cost	<input checked="" type="checkbox"/>	<input type="checkbox"/>
<input checked="" type="checkbox"/>	Promotion Key	<input checked="" type="checkbox"/>	<input type="checkbox"/>
<input checked="" type="checkbox"/>	Revision Number	<input checked="" type="checkbox"/>	<input type="checkbox"/>
<input checked="" type="checkbox"/>	Sales Amount	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
<input checked="" type="checkbox"/>	Sales Territory Key	<input checked="" type="checkbox"/>	<input type="checkbox"/>
<input checked="" type="checkbox"/>	Tax Amt	<input checked="" type="checkbox"/>	<input type="checkbox"/>
<input checked="" type="checkbox"/>	Total Product Cost	<input checked="" type="checkbox"/>	<input type="checkbox"/>

Add Nested Tables   Remove Nested Table

< Back   Next >   Finish >>   Cancel



Mining Structure Mining Models Mining Model Viewer Mining Accuracy Chart Mining Model Prediction

Mining Model: Dim Product 1 Viewer: Microsoft Association Rules Viewer

Rules Itemsets Dependency Network

Minimum probability: 1.00 Filter Rule:

Minimum importance: 0.40 Show: Show attribute name and value

Show long name Maximum rows: 2000

Pr...	Importance	Rule
1.000	0.404	Product Standard Cost >= 1943.1015585792, Revision Number = 1 - 2 -> Sales Amount >= 2607.0361554944
1.000	0.404	Total Product Cost = 729.1263389696 - 1263.002115072 -> Order Quantity = 1 - 2
1.000	0.404	Total Product Cost = 729.1263389696 - 1263.002115072, Product Standard Cost = 744.185430016 - 1230.813227008 -> Sales Amount >= 2607.0361554944
1.000	0.404	Total Product Cost = 729.1263389696 - 1263.002115072, Product Standard Cost = 744.185430016 - 1230.813227008 -> Order Quantity = 1 - 2
1.000	0.404	Total Product Cost = 729.1263389696 - 1263.002115072, Extended Amount = 2169.9258126336 - 2694.4941436928 -> Sales Amount >= 2607.0361554944
1.000	0.404	Total Product Cost = 729.1263389696 - 1263.002115072, Extended Amount = 2169.9258126336 - 2694.4941436928 -> Order Quantity = 1 - 2
1.000	0.404	Total Product Cost = 729.1263389696 - 1263.002115072, Unit Price >= 2610.761242624 -> Sales Amount >= 2607.0361554944
1.000	0.404	Total Product Cost = 729.1263389696 - 1263.002115072, Unit Price >= 2610.761242624 -> Order Quantity = 1 - 2
1.000	0.404	Total Product Cost = 729.1263389696 - 1263.002115072, Tax Amt >= 211.6976865536 -> Sales Amount >= 2607.0361554944
1.000	0.404	Total Product Cost = 729.1263389696 - 1263.002115072, Tax Amt >= 211.6976865536 -> Order Quantity = 1 - 2
1.000	0.404	Total Product Cost = 729.1263389696 - 1263.002115072, Sales Territory Key >= 9 -> Sales Amount >= 2607.0361554944
1.000	0.404	Total Product Cost = 729.1263389696 - 1263.002115072, Sales Territory Key >= 9 -> Order Quantity = 1 - 2
1.000	0.404	Total Product Cost = 729.1263389696 - 1263.002115072, Revision Number = 1 - 2 -> Sales Amount >= 2607.0361554944
1.000	0.404	Total Product Cost = 729.1263389696 - 1263.002115072, Promotion Key = 1 - 2 -> Sales Amount >= 2607.0361554944
1.000	0.404	Total Product Cost = 729.1263389696 - 1263.002115072, Order Quantity = 1 - 2 -> Sales Amount >= 2607.0361554944
1.000	0.404	Total Product Cost = 729.1263389696 - 1263.002115072, Sales Amount >= 2607.0361554944 -> Order Quantity = 1 - 2
1.000	0.404	Total Product Cost = 729.1263389696 - 1263.002115072, Freight >= 65.0225675392 -> Sales Amount >= 2607.0361554944
1.000	0.404	Total Product Cost = 729.1263389696 - 1263.002115072, Fact Internet Sales Count = 1 - 2 -> Sales Amount >= 2607.0361554944
1.000	0.404	Total Product Cost = 729.1263389696 - 1263.002115072, Currency Key >= 98 -> Sales Amount >= 2607.0361554944
1.000	0.404	Total Product Cost = 729.1263389696 - 1263.002115072, Revision Number = 1 - 2 -> Order Quantity = 1 - 2
1.000	0.404	Total Product Cost = 729.1263389696 - 1263.002115072, Promotion Key = 1 - 2 -> Order Quantity = 1 - 2
1.000	0.404	Total Product Cost = 729.1263389696 - 1263.002115072, Freight >= 65.0225675392 -> Order Quantity = 1 - 2
1.000	0.404	Total Product Cost = 729.1263389696 - 1263.002115072, Fact Internet Sales Count = 1 - 2 -> Order Quantity = 1 - 2
1.000	0.404	Total Product Cost = 729.1263389696 - 1263.002115072, Currency Key >= 98 -> Order Quantity = 1 - 2
1.000	0.404	Product Standard Cost = 744.185430016 - 1230.813227008 -> Sales Amount >= 2607.0361554944
1.000	0.404	Product Standard Cost = 744.185430016 - 1230.813227008 -> Order Quantity = 1 - 2
1.000	0.404	Product Standard Cost = 744.185430016 - 1230.813227008, Unit Price >= 2610.761242624 -> Sales Amount >= 2607.0361554944
1.000	0.404	Product Standard Cost = 744.185430016 - 1230.813227008, Unit Price >= 2610.761242624 -> Order Quantity = 1 - 2
1.000	0.404	Product Standard Cost = 744.185430016 - 1230.813227008, Tax Amt >= 211.6976865536 -> Sales Amount >= 2607.0361554944
1.000	0.404	Product Standard Cost = 744.185430016 - 1230.813227008, Tax Amt >= 211.6976865536 -> Order Quantity = 1 - 2
1.000	0.404	Product Standard Cost = 744.185430016 - 1230.813227008, Sales Territory Key >= 9 -> Sales Amount >= 2607.0361554944
1.000	0.404	Product Standard Cost = 744.185430016 - 1230.813227008, Sales Territory Key >= 9 -> Order Quantity = 1 - 2
1.000	0.404	Product Standard Cost = 744.185430016 - 1230.813227008, Revision Number = 1 - 2 -> Sales Amount >= 2607.0361554944
1.000	0.404	Product Standard Cost = 744.185430016 - 1230.813227008, Promotion Key = 1 - 2 -> Sales Amount >= 2607.0361554944
1.000	0.404	Product Standard Cost = 744.185430016 - 1230.813227008, Order Quantity = 1 - 2 -> Sales Amount >= 2607.0361554944
1.000	0.404	Product Standard Cost = 744.185430016 - 1230.813227008, Sales Amount >= 2607.0361554944 -> Order Quantity = 1 - 2
1.000	0.404	Product Standard Cost = 744.185430016 - 1230.813227008, Freight >= 65.0225675392 -> Sales Amount >= 2607.0361554944

Rules: 222

## Przykład - grupowanie

- Na podstawie atrybutów opisujących klienta predykujemy OrderQty

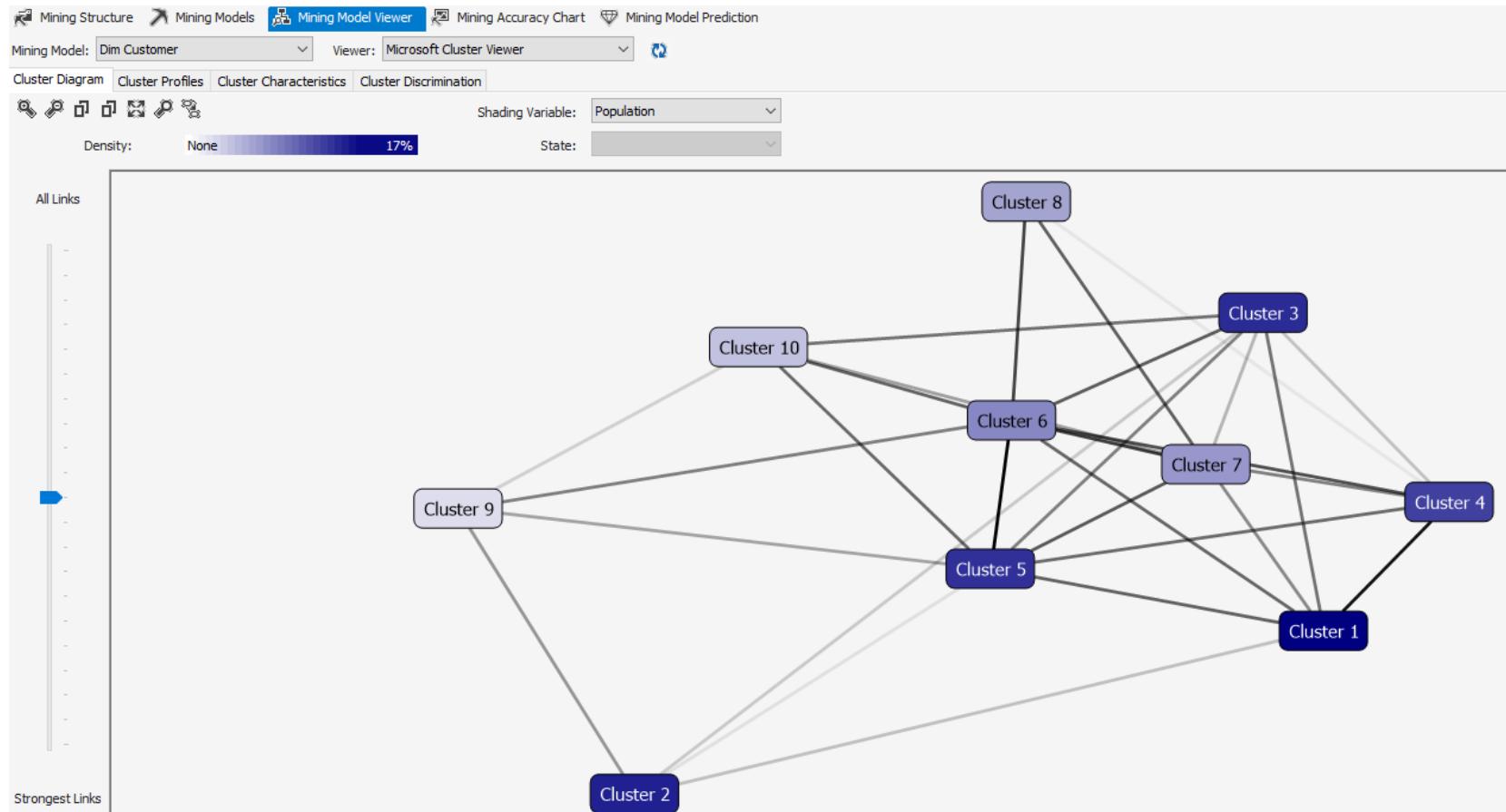
**Specify Mining Model Column Usage**  
Specify the usage of mining model columns and optionally add nested tables.



Mining model structure:

	Tables/Columns	<input type="checkbox"/> Input	<input type="checkbox"/> Predict...
<input checked="" type="checkbox"/>	Customer Key	<input checked="" type="checkbox"/>	<input type="checkbox"/>
<input checked="" type="checkbox"/>	Commute Distance	<input checked="" type="checkbox"/>	<input type="checkbox"/>
<input checked="" type="checkbox"/>	English Education	<input checked="" type="checkbox"/>	<input type="checkbox"/>
<input checked="" type="checkbox"/>	Gender	<input checked="" type="checkbox"/>	<input type="checkbox"/>
<input checked="" type="checkbox"/>	Marital Status	<input checked="" type="checkbox"/>	<input type="checkbox"/>
<input checked="" type="checkbox"/>	Title	<input checked="" type="checkbox"/>	<input type="checkbox"/>
<input checked="" type="checkbox"/>	Total Children	<input checked="" type="checkbox"/>	<input type="checkbox"/>
<input checked="" type="checkbox"/>	Yearly Income	<input checked="" type="checkbox"/>	<input type="checkbox"/>
<input checked="" type="checkbox"/>	Order Quantity	<input type="checkbox"/>	<input checked="" type="checkbox"/>

# Przykład - grupowanie





„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”

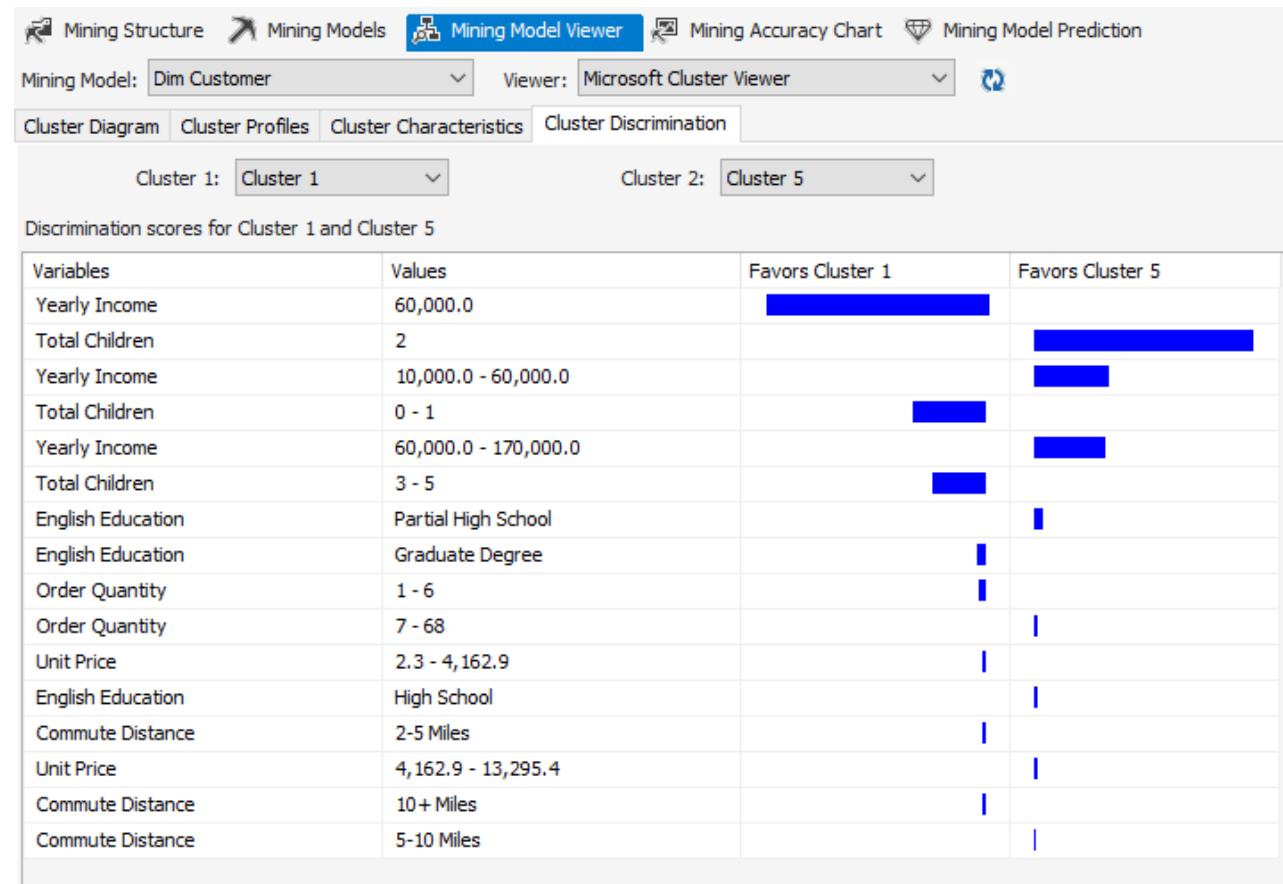
# Przykład - grupowanie





„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”

# Przykład - grupowanie





Fundusze  
Europejskie  
Wiedza Edukacja Rozwój



Politechnika Wrocławska

Unia Europejska  
Europejski Fundusz Społeczny



*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Hurtownie danych

Dziękuję za uwagę

dr inż. Marcin Maleszka



Fundusze  
Europejskie  
Wiedza Edukacja Rozwój



Politechnika Wrocławskiego

Unia Europejska  
Europejski Fundusz Społeczny



*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Hurtownie danych

**Podstawy wizualizacji danych**

**dr inż. Marcin Maleszka**



„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”

- Tabela
- Interpretacja?

## Sposób prezentacji danych

	K1	K2	K3	K4	K5	Razem
<b>A1</b>	26 573	13 009	19 177	26 574	9 656	<b>94 989</b>
<b>A2</b>	24 574	10 394	9 756	13 299	3 464	<b>61 487</b>
<b>A3</b>	12 834	11 062	10 107	24 727	8 448	<b>67 178</b>
<b>A4</b>	23 071	5 757	16 048	16 622	8 820	<b>70 318</b>
<b>B1</b>	12 389	12 086	18 732	19 761	9 219	<b>72 187</b>
<b>B2</b>	21 947	6 307	11 418	28 864	3 414	<b>71 950</b>
<b>B3</b>	9 873	10 663	17 500	20 081	4 796	<b>62 913</b>
<b>B4</b>	12 104	5 833	14 293	28 291	8 805	<b>69 326</b>
<b>C1</b>	20 008	10 768	17 403	26 808	4 748	<b>79 735</b>
<b>C2</b>	16 299	11 979	12 843	28 541	4 013	<b>73 675</b>
<b>C3</b>	24 337	6 726	10 752	15 075	8 861	<b>65 751</b>
<b>C4</b>	12 936	11 635	15 914	23 313	5 534	<b>69 332</b>
<b>Razem</b>	<b>216 945</b>	<b>116 219</b>	<b>173 943</b>	<b>271 956</b>	<b>79 778</b>	<b>858 841</b>



„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”

- Tabela
- Interpretacja?

## Sposób prezentacji danych

	K1	K2	K3	K4	K5	Razem
<b>A1</b>	26 573	13 009	19 177	26 574	9 656	<b>94 989</b>
<b>A2</b>	24 574	10 394	9 756	13 299	3 464	<b>61 487</b>
<b>A3</b>	12 834	11 062	10 107	24 727	8 448	<b>67 178</b>
<b>A4</b>	23 071	5 757	16 048	16 622	8 820	<b>70 318</b>
<b>B1</b>	12 389	12 086	18 732	19 761	9 219	<b>72 187</b>
<b>B2</b>	21 947	6 307	11 418	28 864	3 414	<b>71 950</b>
<b>B3</b>	9 873	10 663	17 500	20 081	4 796	<b>62 913</b>
<b>B4</b>	12 104	5 833	14 293	28 291	8 805	<b>69 326</b>
<b>C1</b>	20 008	10 768	17 403	26 808	4 748	<b>79 735</b>
<b>C2</b>	16 299	11 979	12 843	28 541	4 013	<b>73 675</b>
<b>C3</b>	24 337	6 726	10 752	15 075	8 861	<b>65 751</b>
<b>C4</b>	12 936	11 635	15 914	23 313	5 534	<b>69 332</b>
<b>Razem</b>	<b>216 945</b>	<b>116 219</b>	<b>173 943</b>	<b>271 956</b>	<b>79 778</b>	<b>858 841</b>

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Wizualizacja

- Ważna część zrozumienia informacji
- Ułatwienie wyszukiwania informacji i podejmowania decyzji
- Wsparcie analizy większych zestawów danych
- Wykrycie dodatkowych zależności
- Zmniejszenie wysiłku niezbędnego do przetwarzania informacji
- Ułatwienie zapamiętania danych

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Najważniejsze aspekty

- Właściwości wizualne:
  - kształt, kolor, rozmiar, orientacja, pozycja, czytelność
  - forma wizualna
  - elementy graficzne
  - wskazówki wizualne
  - formatowanie warunkowe
  - infografiki, schematy, wykresy

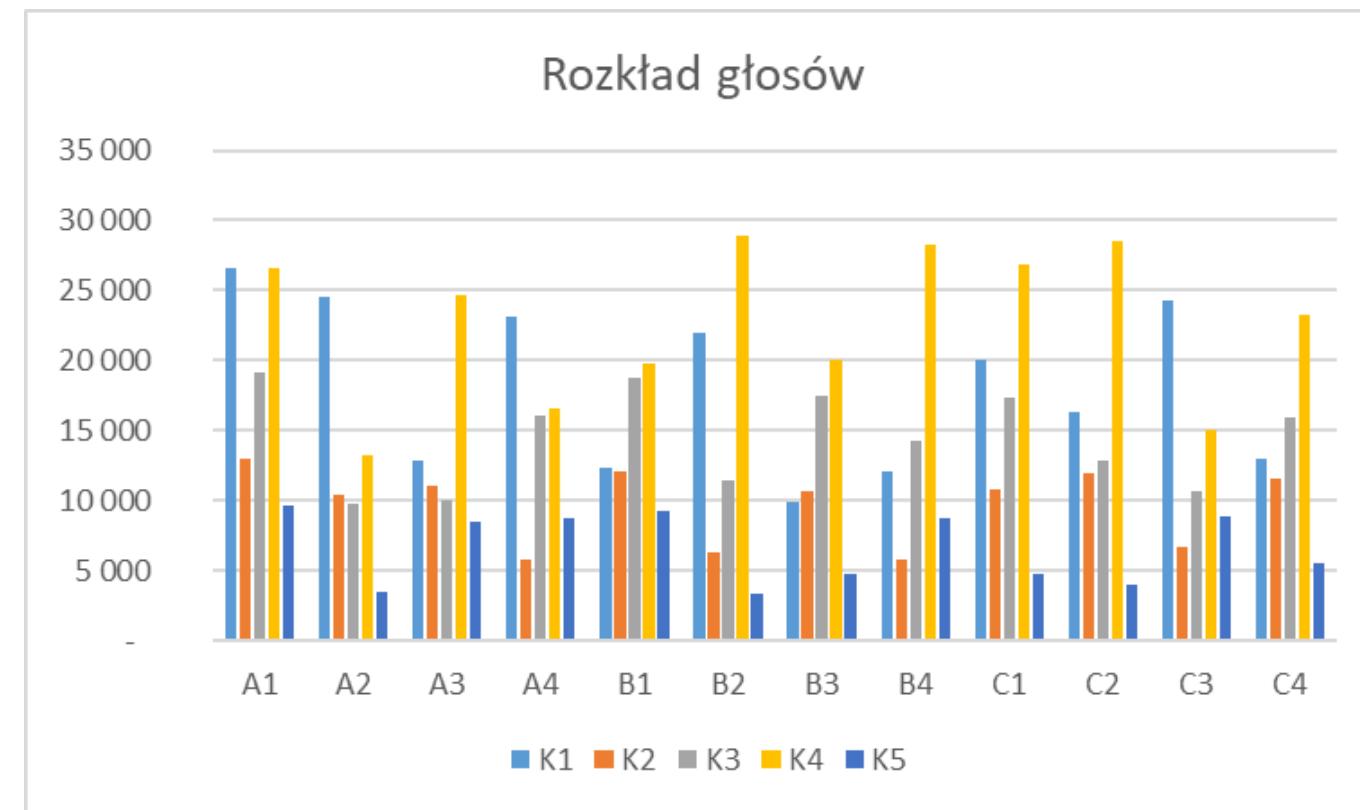
# Najważniejsze aspekty

1. Określ cel wykresu – co chcemy pokazać?
2. Określ, z czym porównujemy
  1. Procent całości
  2. Ranking
  3. Dynamika zmian w czasie
  4. Rozkład częstości (histogram)
  5. Korelacje pomiędzy zmiennymi
3. Przygotuj wykres
4. Sformatuj wykres

„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”

## Sposób prezentacji danych

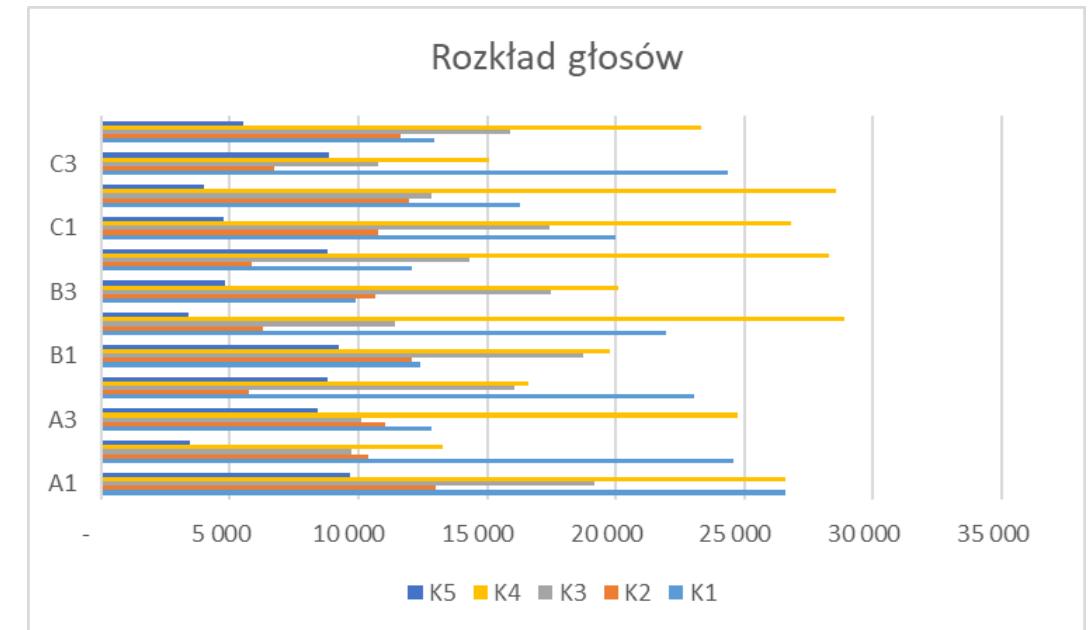
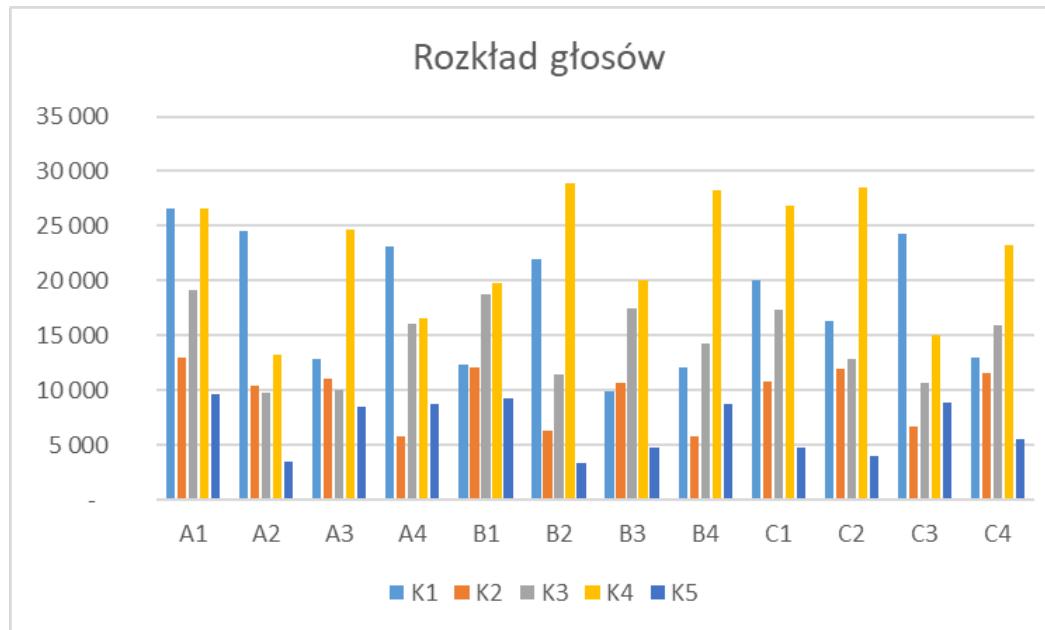
- Wykres
- Interpretacja?



„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”

# Wykres kolumnowy / słupkowy

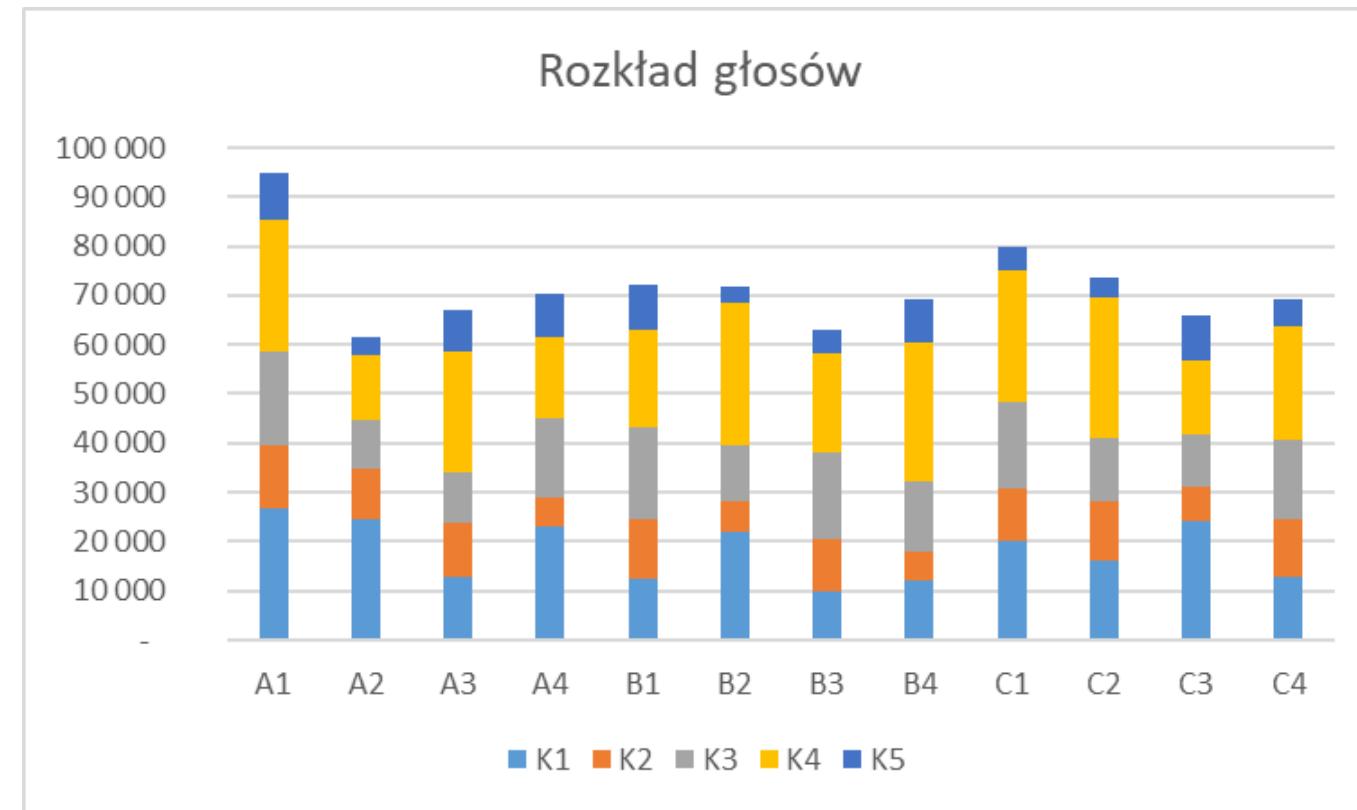
- Zalety?
- Wady?



„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”

## Sposób prezentacji danych

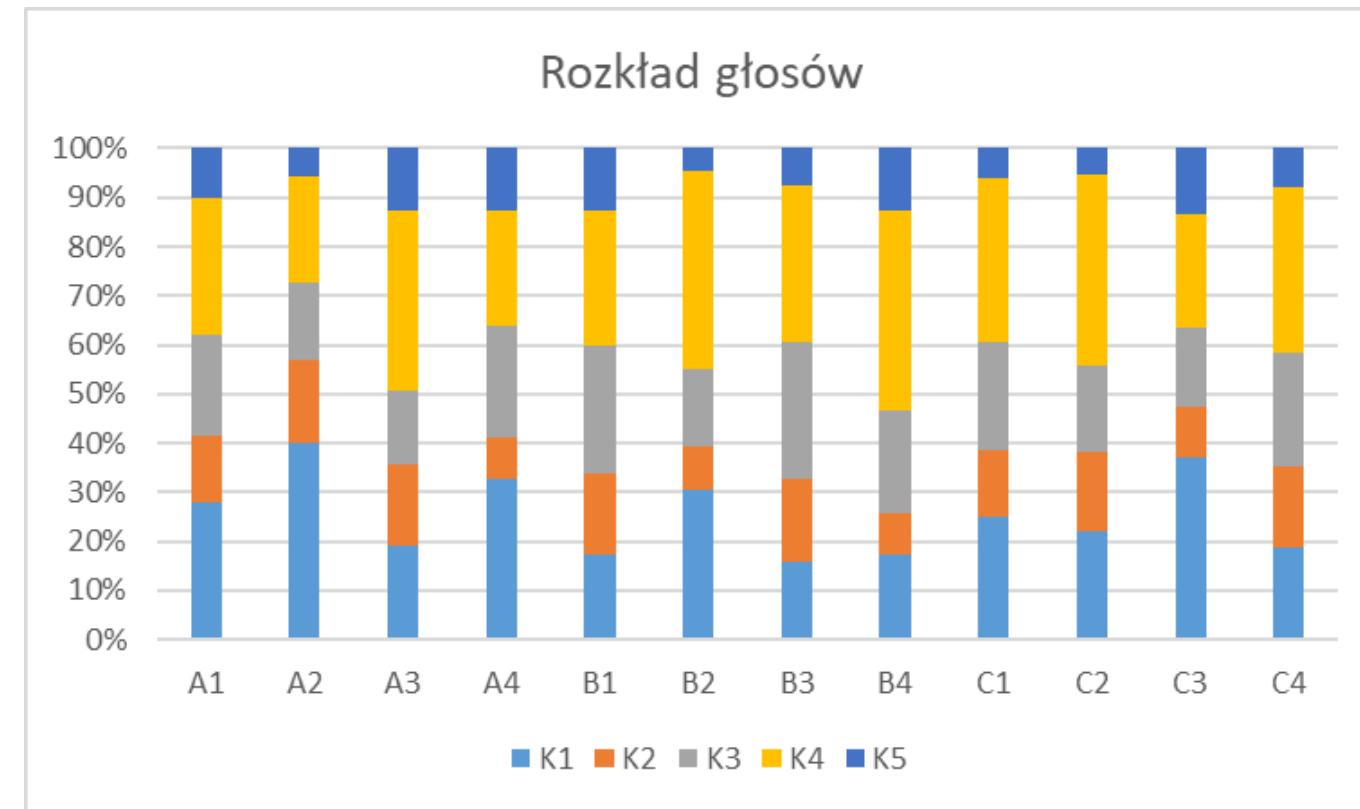
- Wykres
- Interpretacja?



„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”

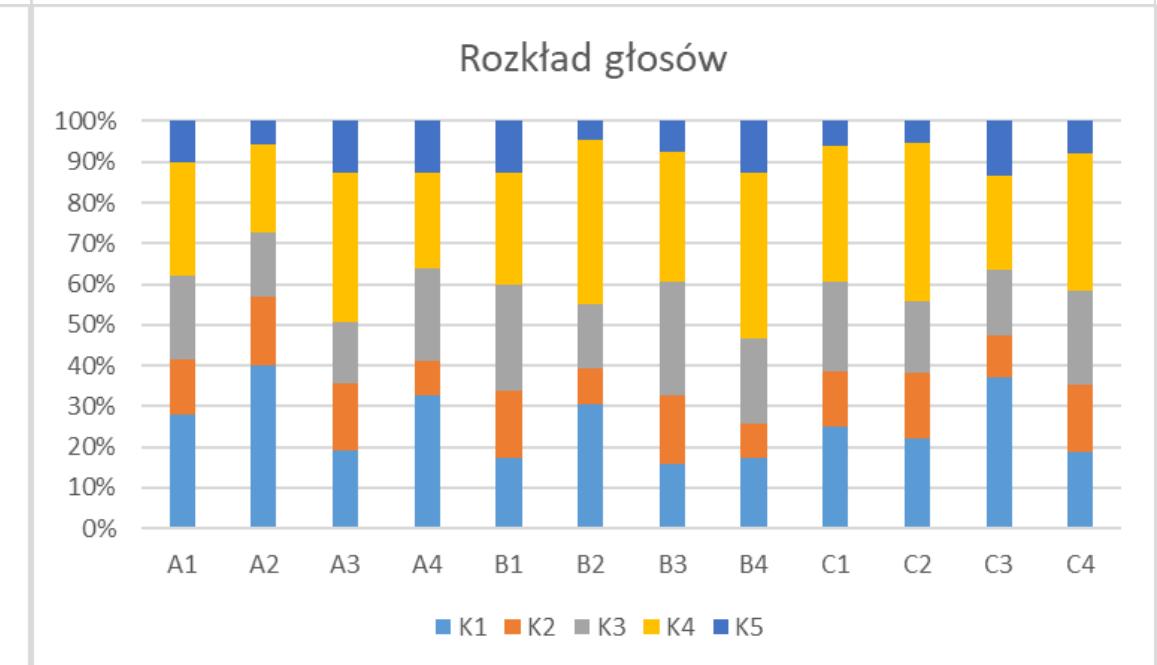
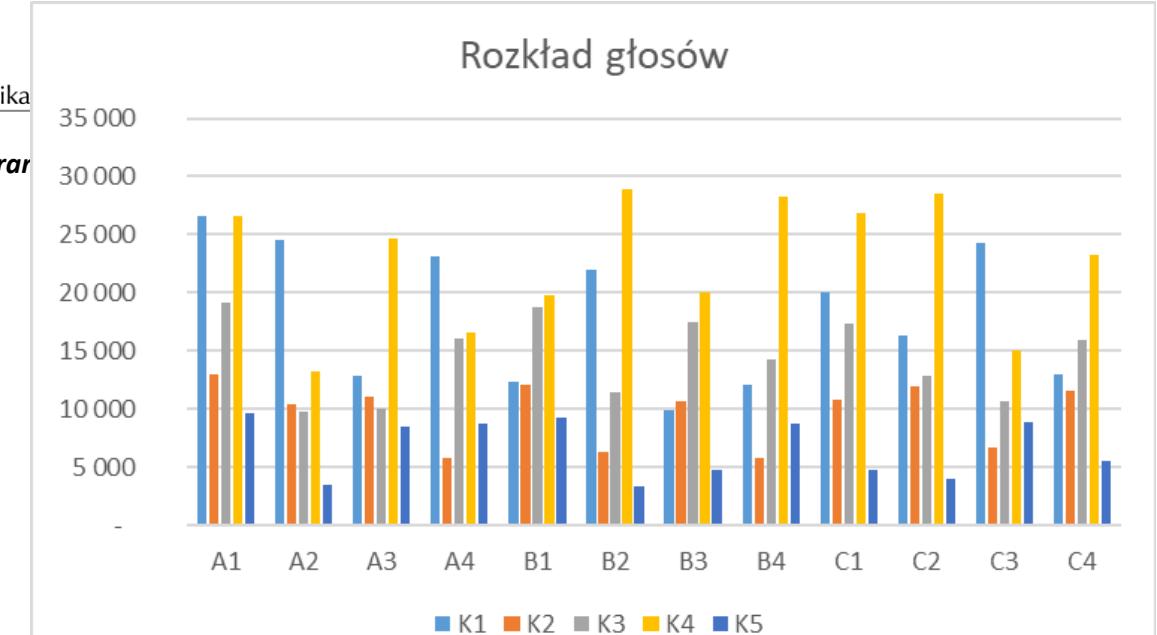
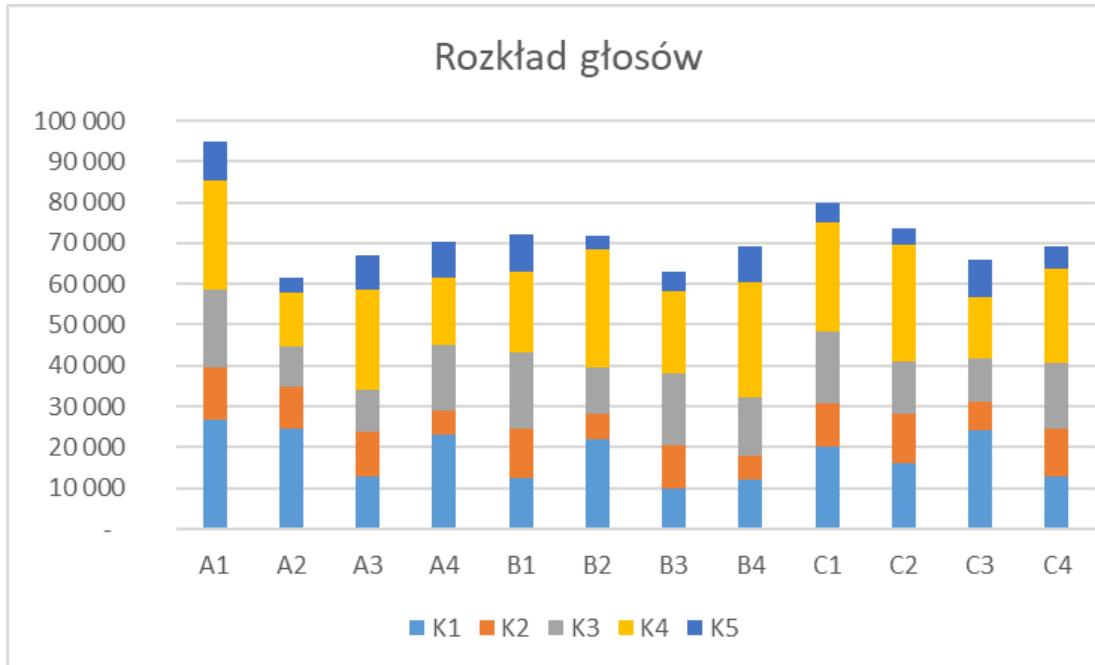
## Sposób prezentacji danych

- Wykres
- Interpretacja?



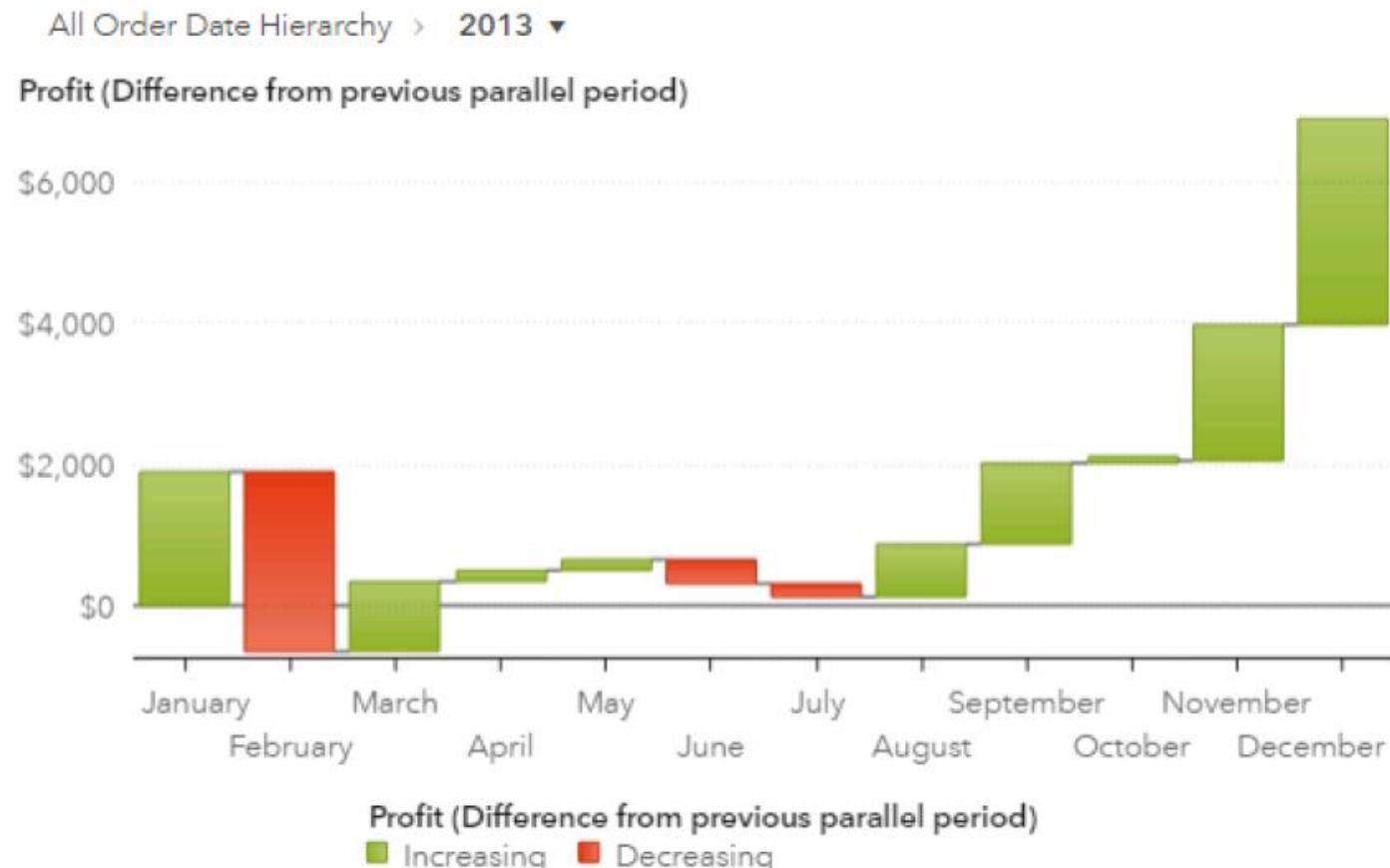


# Wskaż różnice



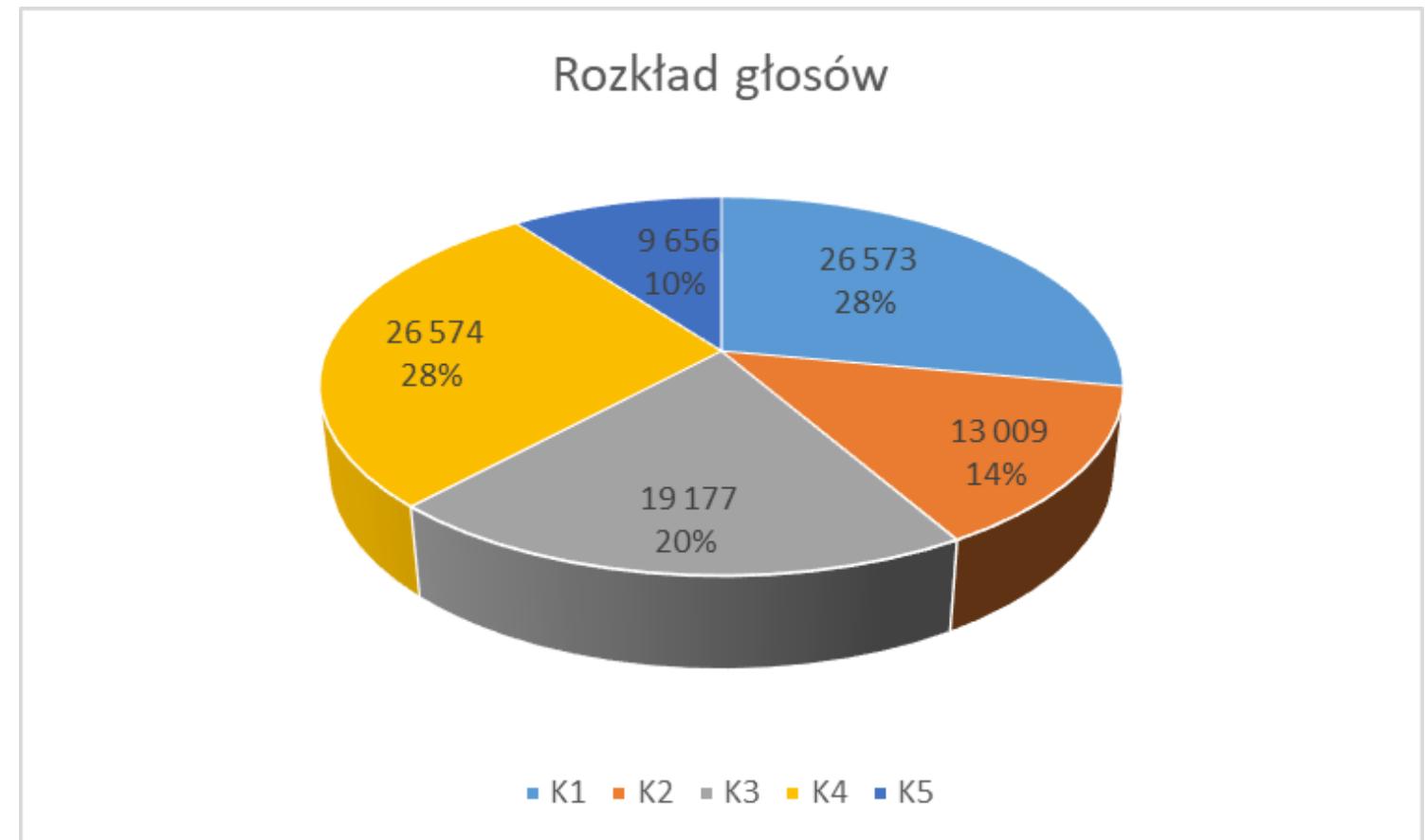
„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”

## Przykład – alternatywa dla wykresu słupkowego



- Zalety?
- Wady?
- Dlaczego nie wystarcza do analiz w hurtowni danych?

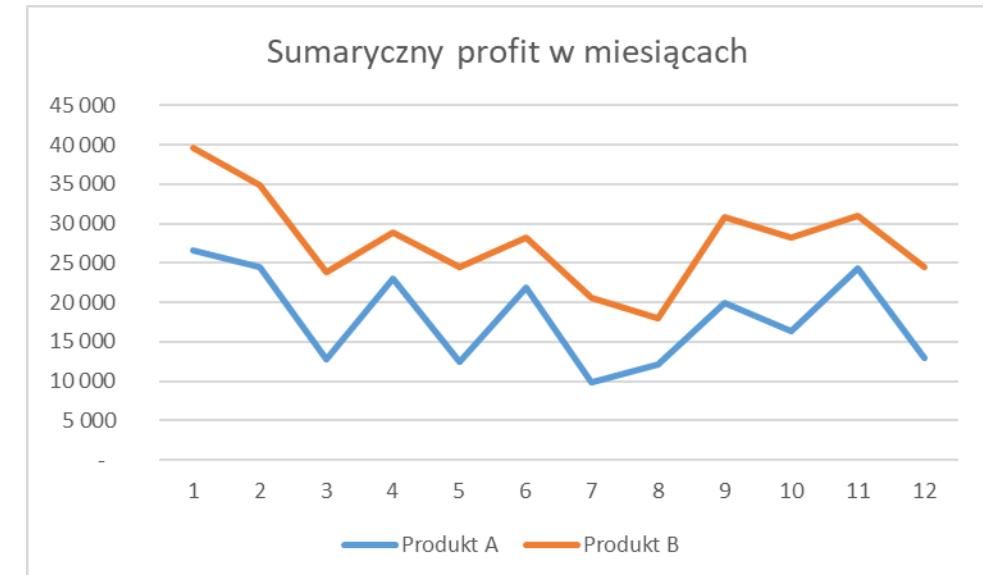
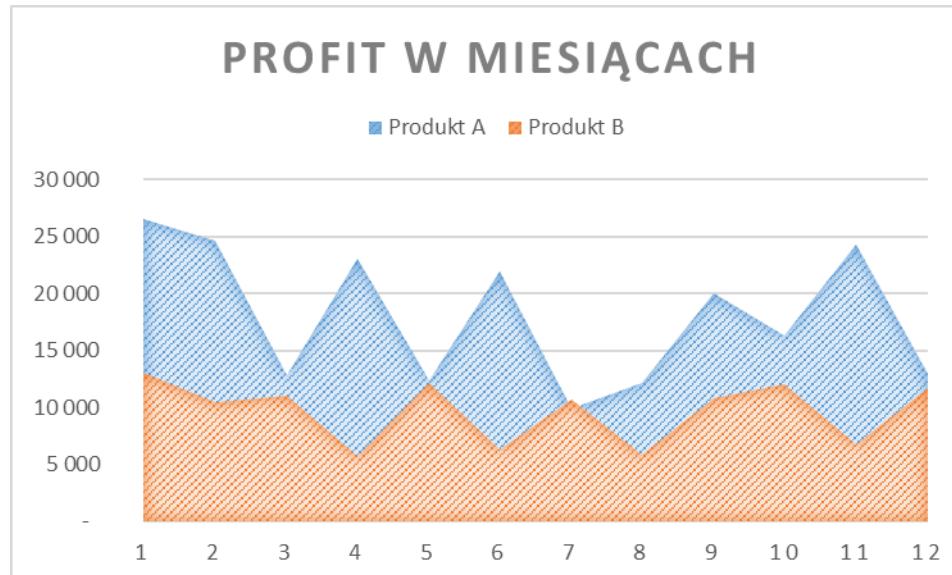
## Wykres kołowy



*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

## Wykres liniowy

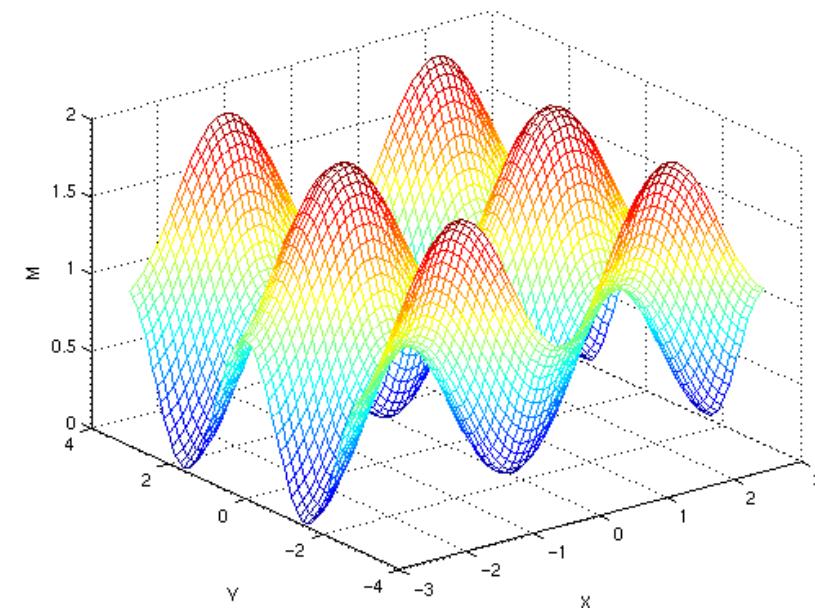
- Kiedy stosować?
- Zalety?
- Wady?



*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Typy wykresów

- Kolumnowy/słupkowy
- Liniowy
- Kołowy/pierścieniowy
- Punktowy
- Giełdowy
- Radarowy
- Histogram
- Boxplot
- Kaskadowy
- Hybrydowy





Fundusze  
Europejskie  
Wiedza Edukacja Rozwój



Politechnika Wrocławskiego

Unia Europejska  
Europejski Fundusz Społeczny



*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

## Przykładowe błędy

- Najczęstsze błędy przy prezentacji wykresów:
  - brak czytelności
  - niewłaściwy dobór typu wykresu
  - brak legendy, opisu osi
  - niewłaściwie dobrana skala / zakres wartości na osiach



Fundusze  
Europejskie  
Wiedza Edukacja Rozwój



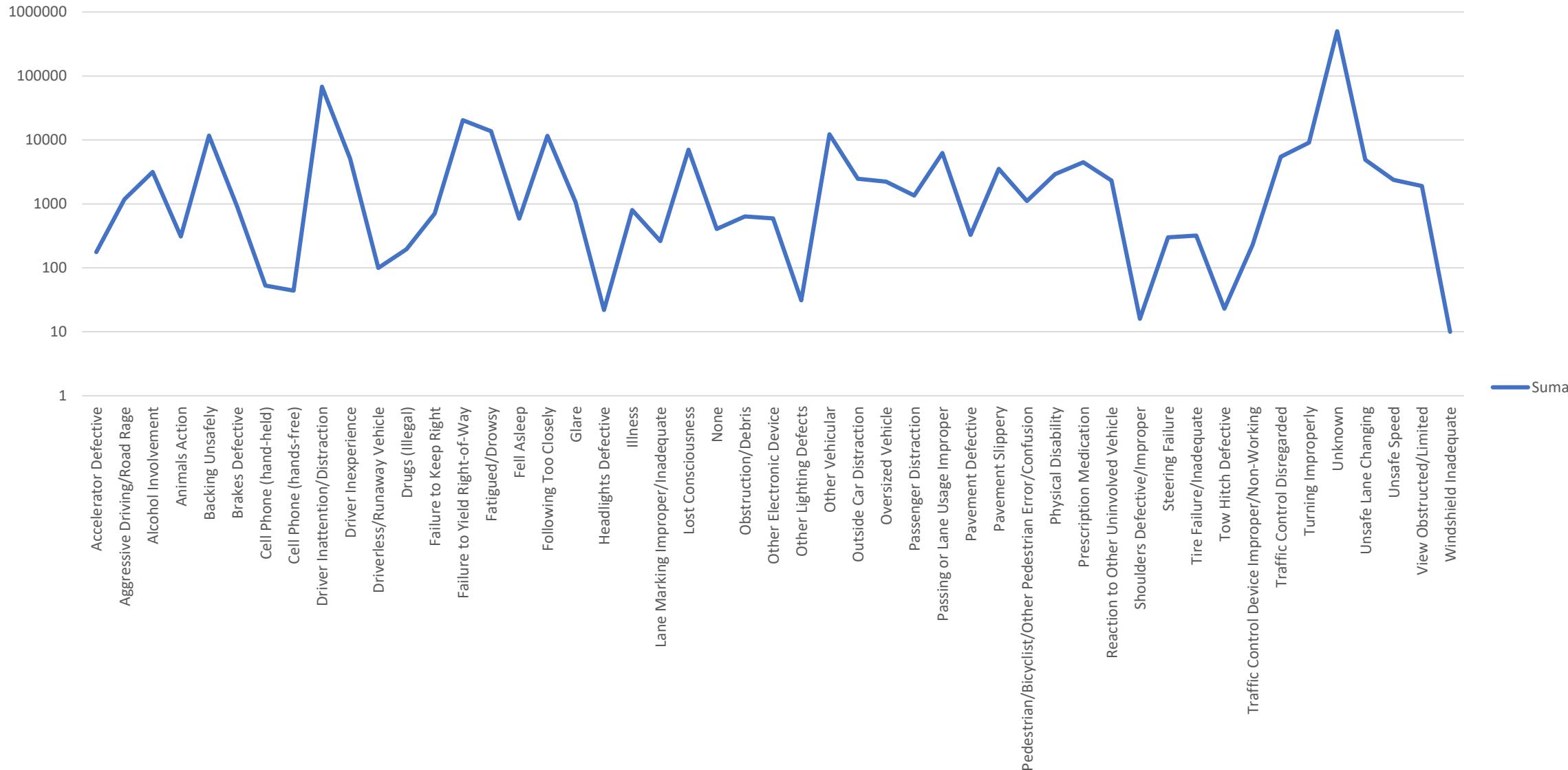
Politechnika Wrocławskiego



Unia Europejska  
Europejski Fundusz Społeczny

**„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”**

Suma





Fundusze  
Europejskie  
Wiedza Edukacja Rozwój



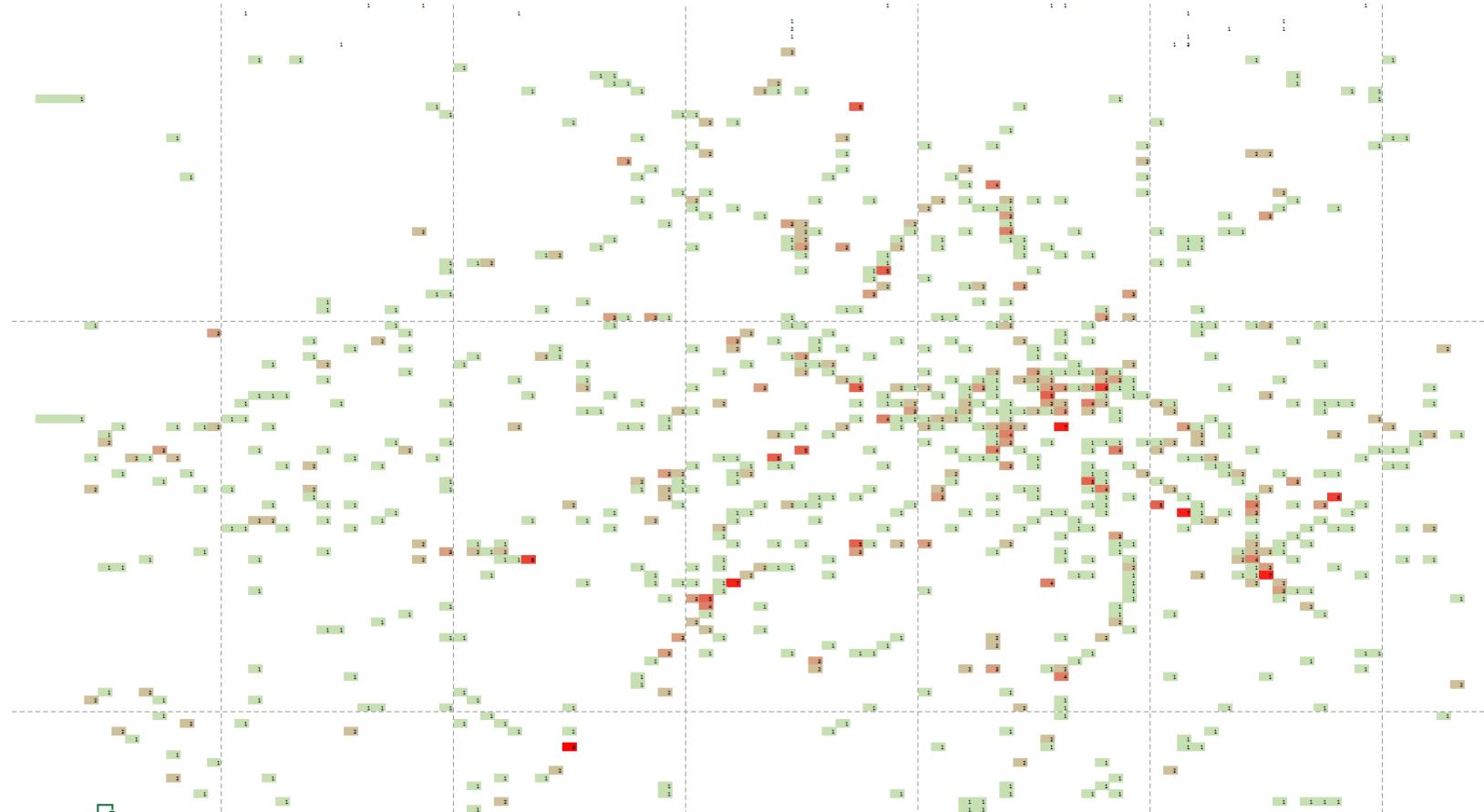
Politechnika Wrocławska

Unia Europejska  
Europejski Fundusz Społeczny



„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”

# Rozmieszczenie wypadków





Fundusze  
Europejskie  
Wiedza Edukacja Rozwój



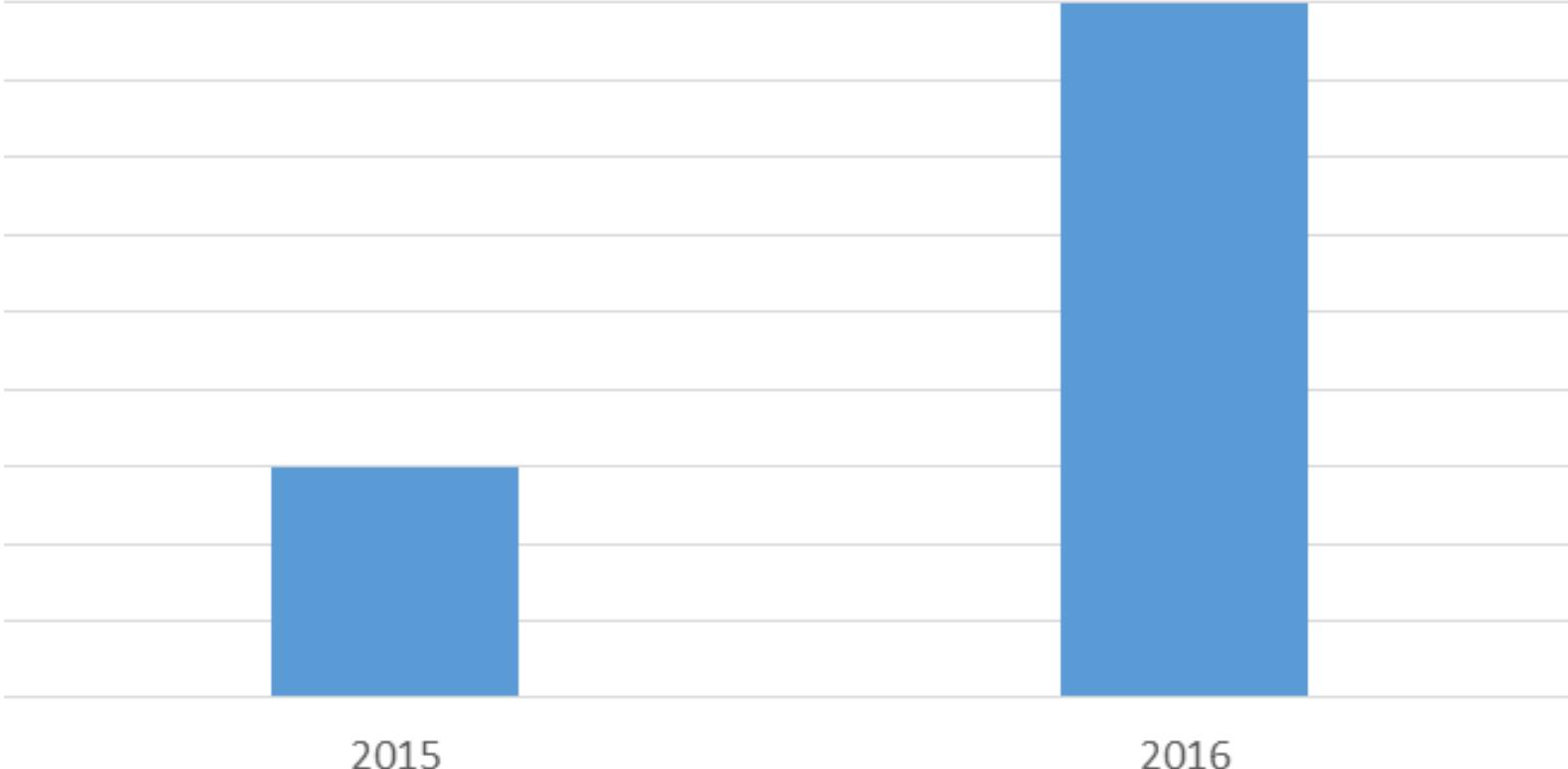
Politechnika Wrocławska

Unia Europejska  
Europejski Fundusz Społeczny



*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

## Zysk





Fundusze  
Europejskie  
Wiedza Edukacja Rozwój

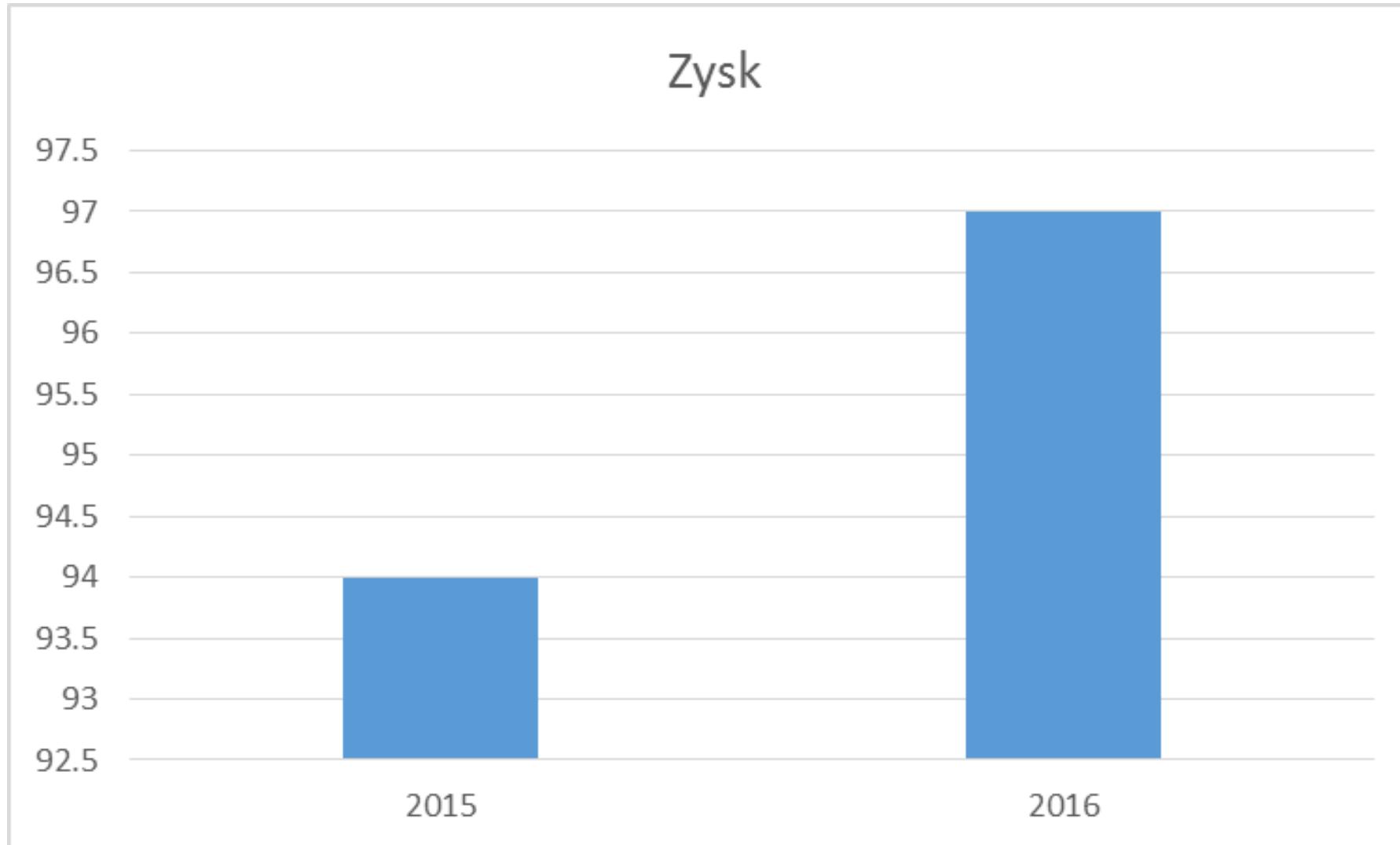


Politechnika Wrocławska

Unia Europejska  
Europejski Fundusz Społeczny

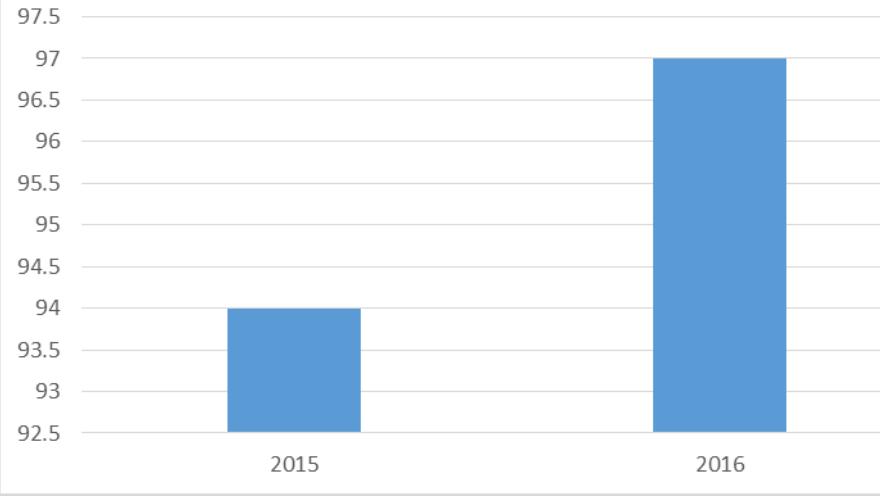


*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

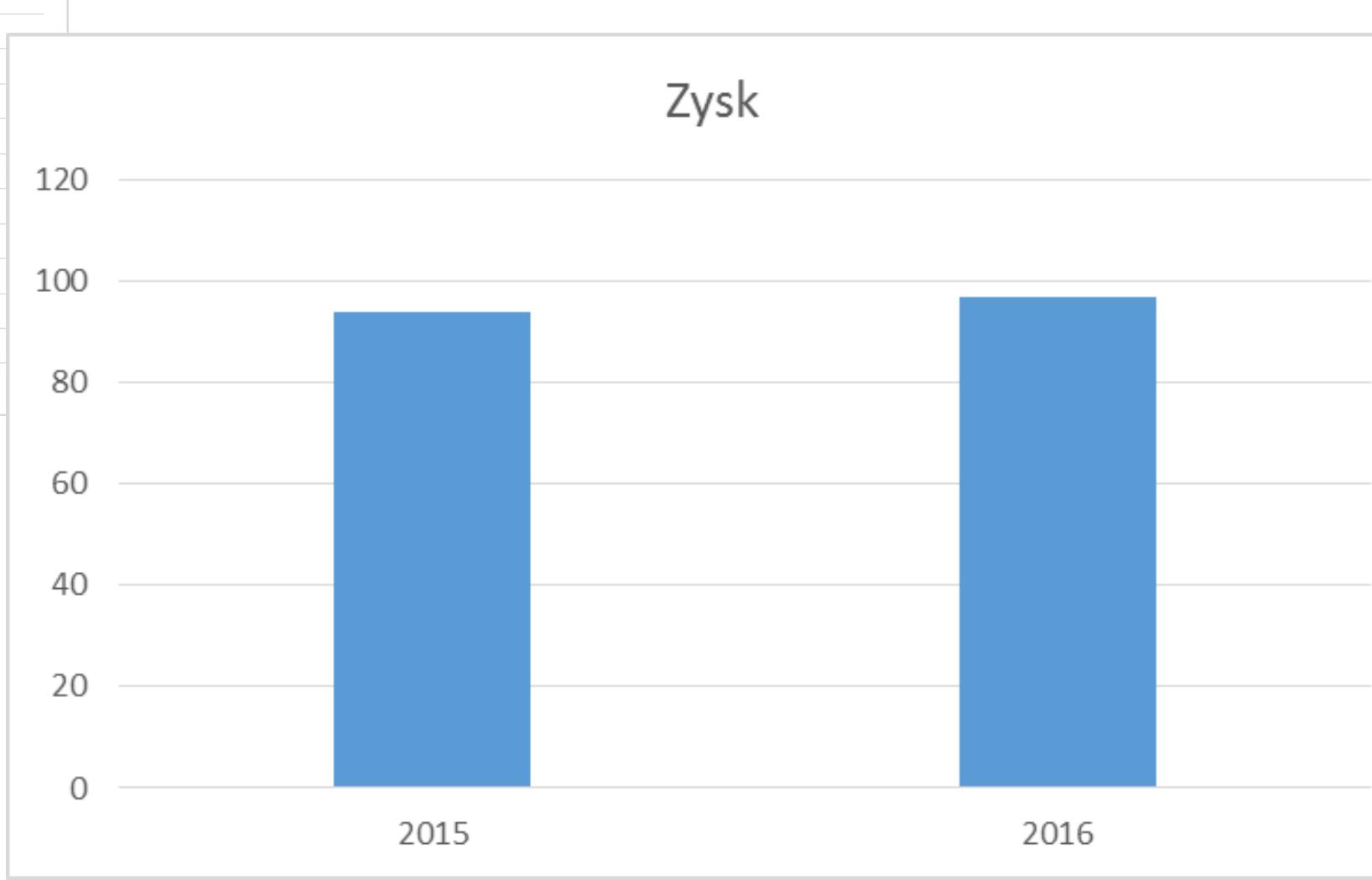


**„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”**

Zysk

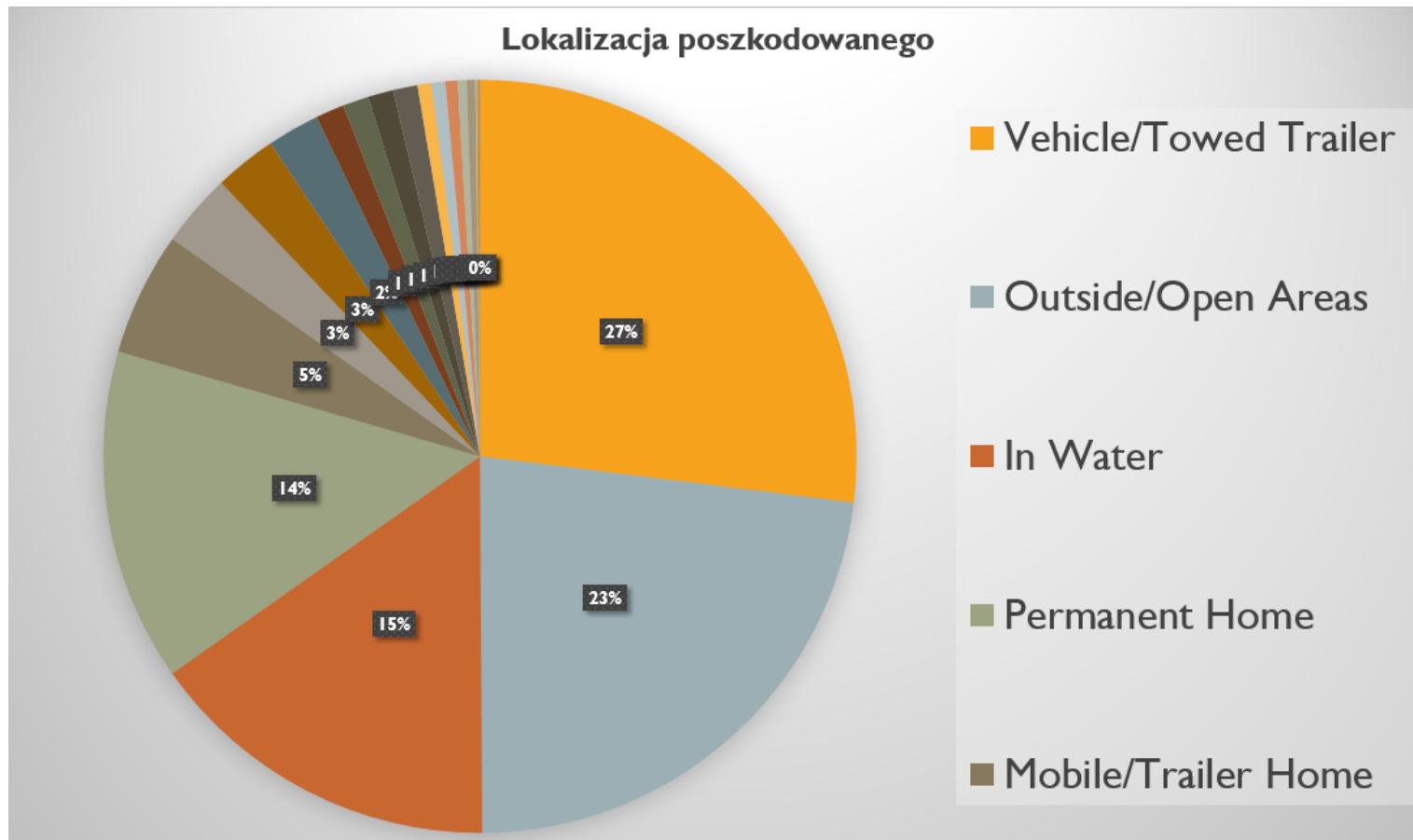


Zysk



„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”

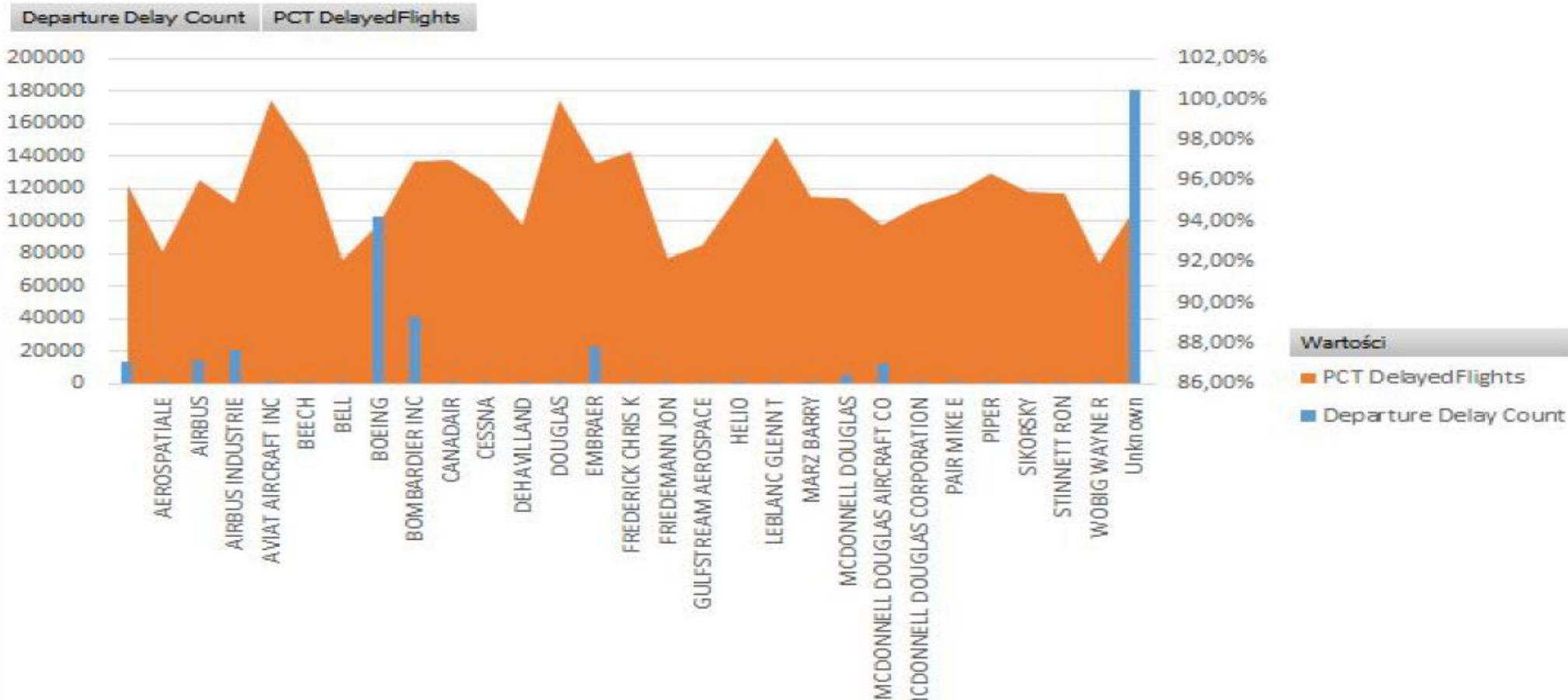
## Zjawiska pogodowe





„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”

# Opóźnienia lotów





Fundusze  
Europejskie  
Wiedza Edukacja Rozwój



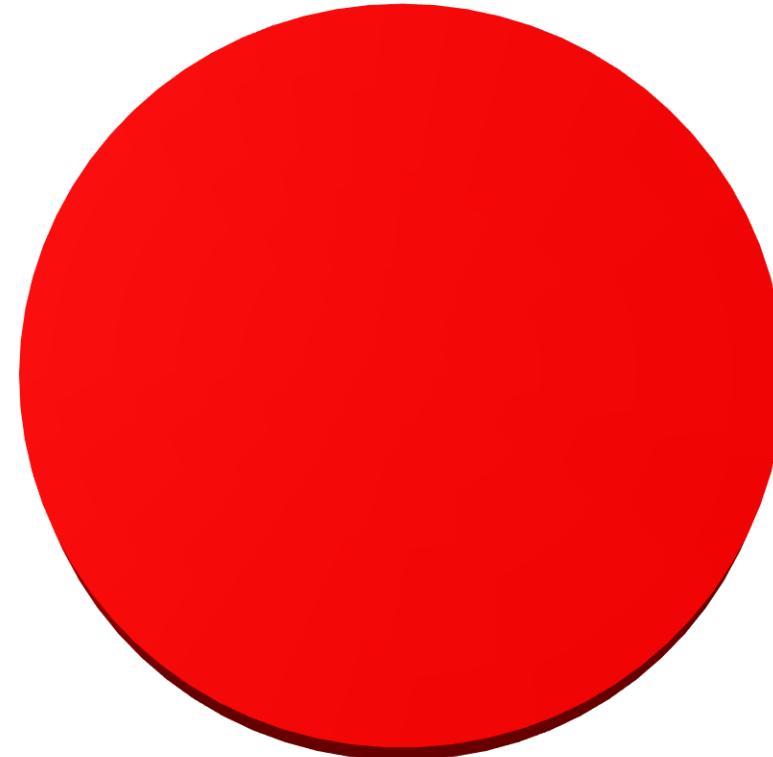
Politechnika Wrocławskiego

Unia Europejska  
Europejski Fundusz Społeczny



*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Źródło pożaru



■ Ogień



Fundusze  
Europejskie  
Wiedza Edukacja Rozwój



Politechnika Wrocławska

Unia Europejska  
Europejski Fundusz Społeczny



*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Hurtownie danych

Dziękuję za uwagę

dr inż. Marcin Maleszka



Fundusze  
Europejskie  
Wiedza Edukacja Rozwój



Politechnika Wrocławska

Unia Europejska  
Europejski Fundusz Społeczny



*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Hurtownie danych

**Podstawy projektowania hurtowni danych**

**dr inż. Marcin Maleszka**

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Business Intelligence

## analityka biznesowa

- Proces przekształcania:
  - danych w informacje
  - informacji w wiedzę
- Zalety hurtowni danych:
  - zwiększenie konkurencyjności przedsiębiorstwa
  - uwolnienie systemów transakcyjnych od tworzenia raportów
  - umożliwienie równoczesnego korzystania z różnych systemów BI

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Systemy BI

- generuje raporty i wylicza kluczowe wskaźniki PKI
- na podstawie raportów stawia się hipotezy
- weryfikacja hipotez na podstawie szczegółowych analiz (OLAP, data mining)
  
- Odmiany/Przykłady BI:
  - EIS – Executive Information Systems
  - DSS – Decision Support Systems
  - MIS – Management Information Systems
  - GIS – Geographic Information Systems



„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”

# Wykorzystanie systemów BI

- menadżerowie niższych szczebli – BAM (Business Activity Monitoring)
- kokpit menadżera
- panel nawigacyjny
- dashboard



# Odkrywanie wiedzy z baz danych

## Knowlegde Discovery in Databases

- wykorzystanie narzędzi informatycznych i dużych zbiorów danych (hurtowni) do znajdowania ukrytych dla człowieka prawidłowości
- stosowane rozwiązania:
  - wizualizacje na wykresach
  - metody statystyczne
  - sieci neuronowe
  - metody uczenia maszynowego
  - metody ewolucyjne
  - logika rozmyta
  - zbiory przybliżone

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Proces odkrywania wiedzy z danych

## Knowlegde Discovery in Databases

- zrozumienie dziedziny problemu
- budowa roboczego zbioru danych
- integracja danych: czyszczenie, przekształcanie i redukcja danych
- eksploracja danych

# Zastosowania hurtowni danych

## Przykład 1: controlling i analizy ekonomiczna

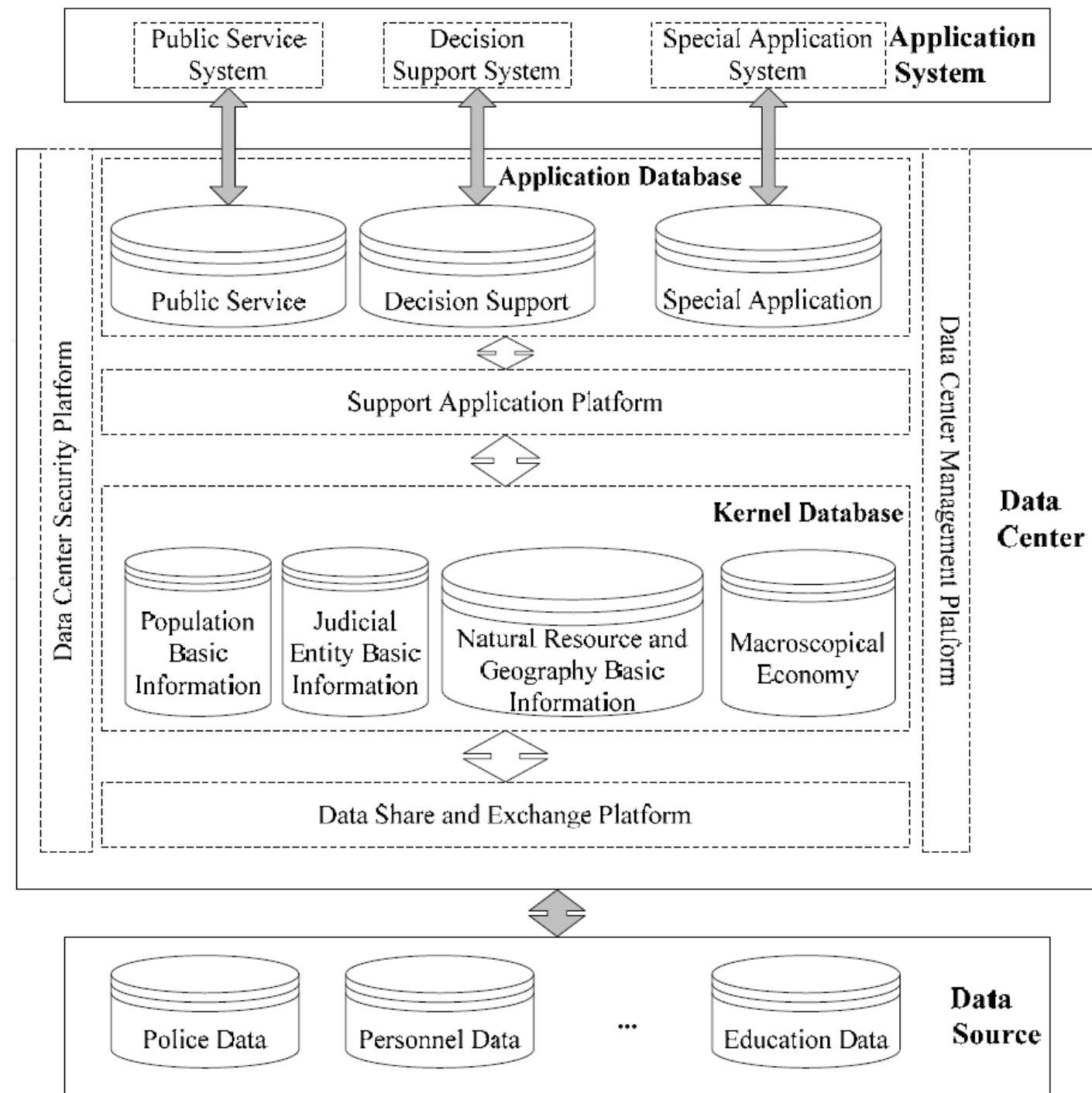
- przed wdrożeniem hurtowni:
  - jeden pracownik przygotowujący raporty w arkuszu kalkulacyjnym
- po wdrożeniu hurtowni:
  - więcej gromadzonych danych
  - 90% raportów mógł wygenerować dowolny użytkownik

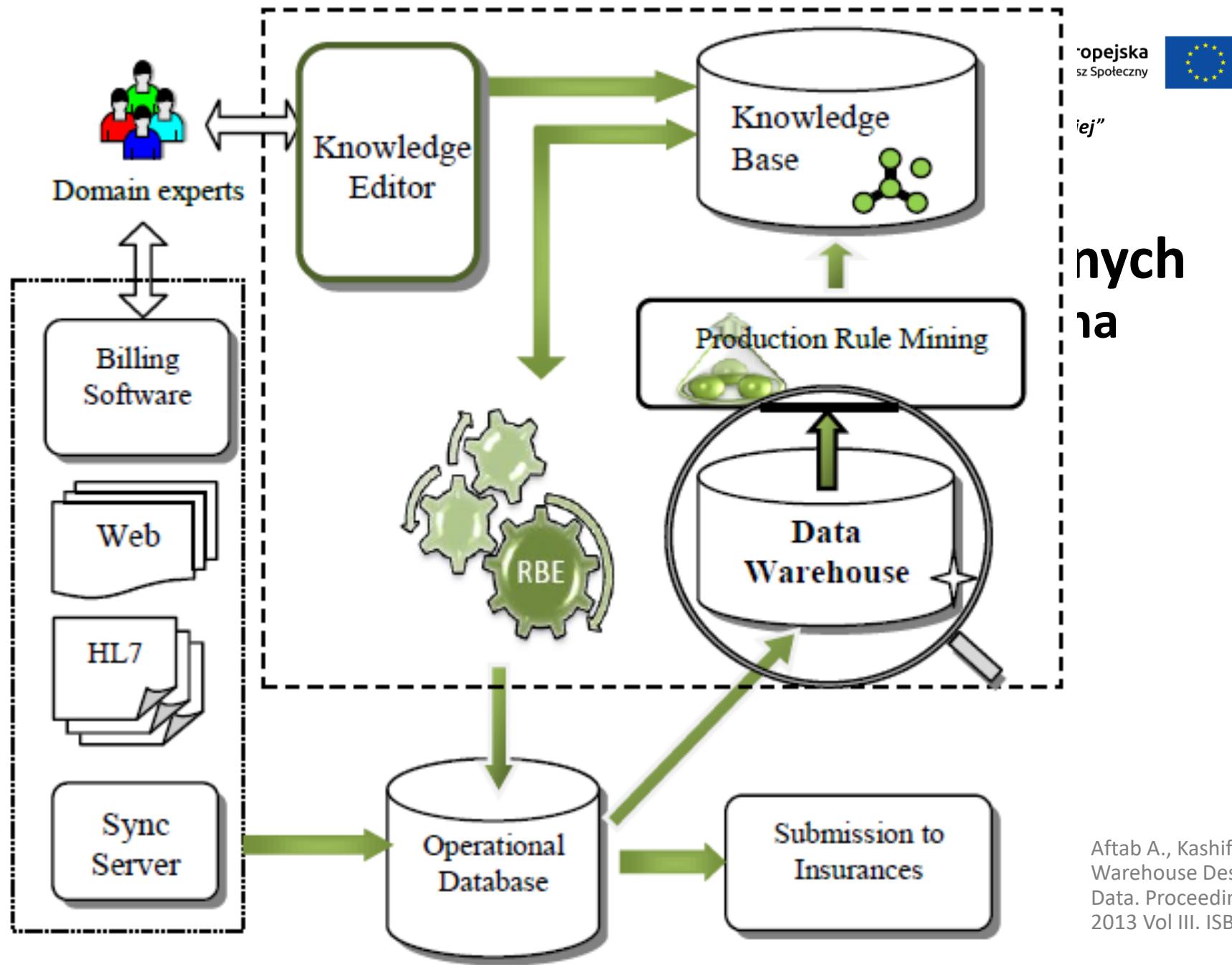


*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

## Zastosowania hurtowni danych Przykład 2: e-government

- usprawnienie działania urzędów w Nanhai
  - Government-to-Citizen (G2C)
  - Government-to-Business (G2B)
  - Government-to-Employee (G2E)
  - Government-to-Government (G2G)
- hurtownia jest podstawą do analizy danych zgromadzonych w centrum informacji





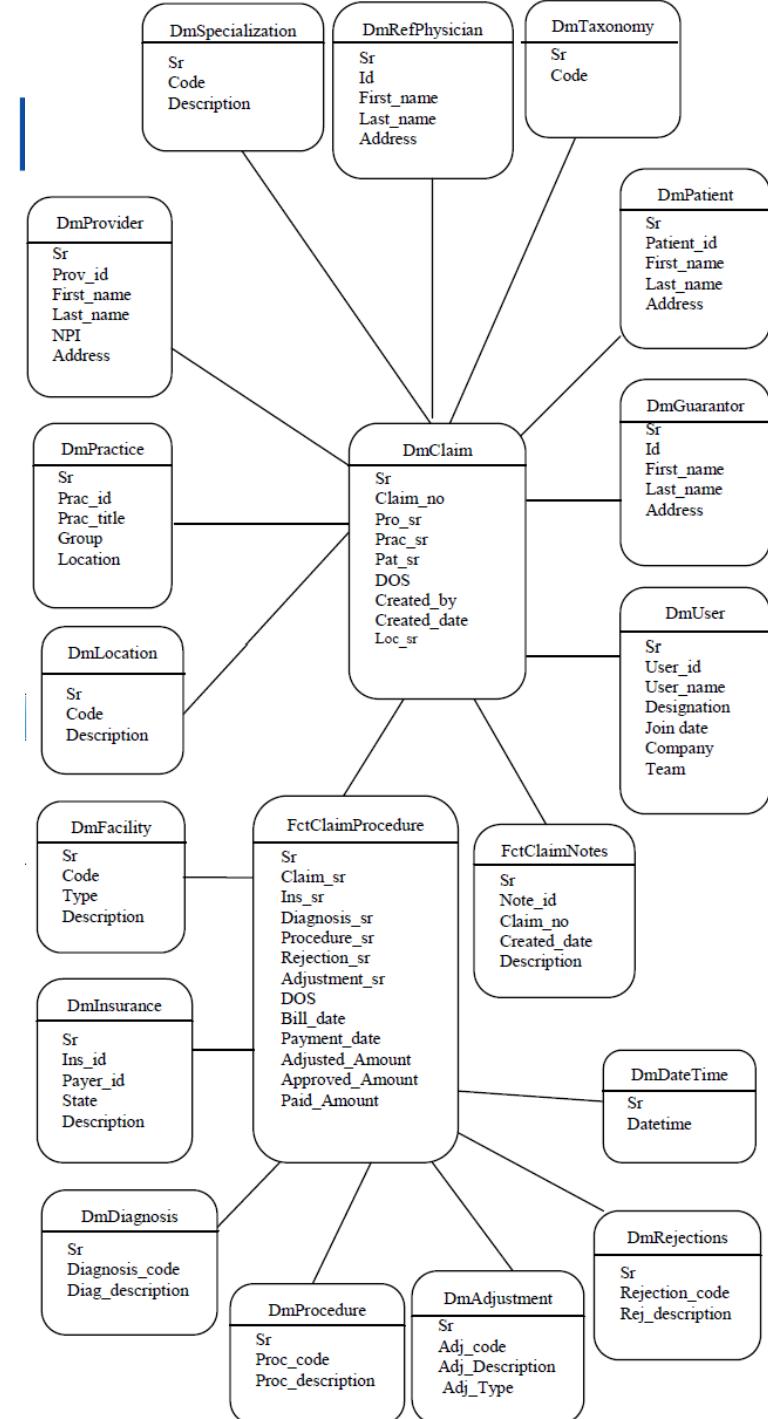
Aftab A., Kashif Z., Abdul B. S., Umair A. (2013): Data Warehouse Design For Knowledge Discovery From Healthcare Data. Proceedings of the World Congress on Engineering 2013 Vol III. ISBN: 978-988-19252-9-9



„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”

## Zastosowania hurtowni danych Przykład 3: opieka zdrowotna

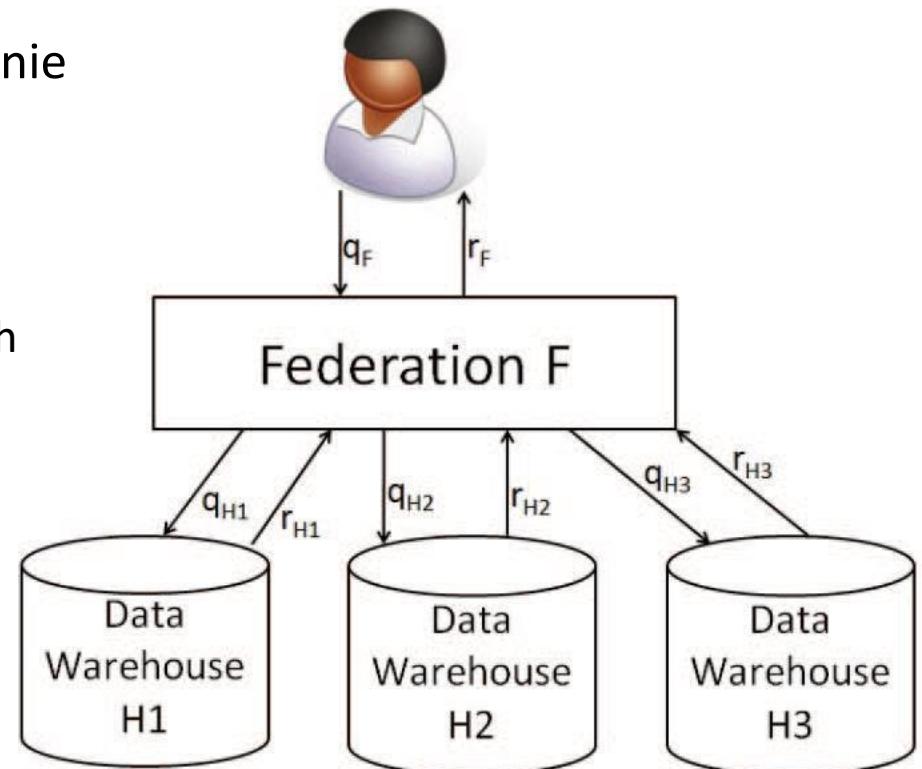
- wykorzystanie rule-based expert system do poprawy jakości świadczonych usług
- dane kliniczne z University of Virginia:
  - diagnozy poprawne i błędne
- poprawa jakości usług:
  - z 5.6% błędnych diagnoz do 3.04%



## Zastosowania hurtowni danych

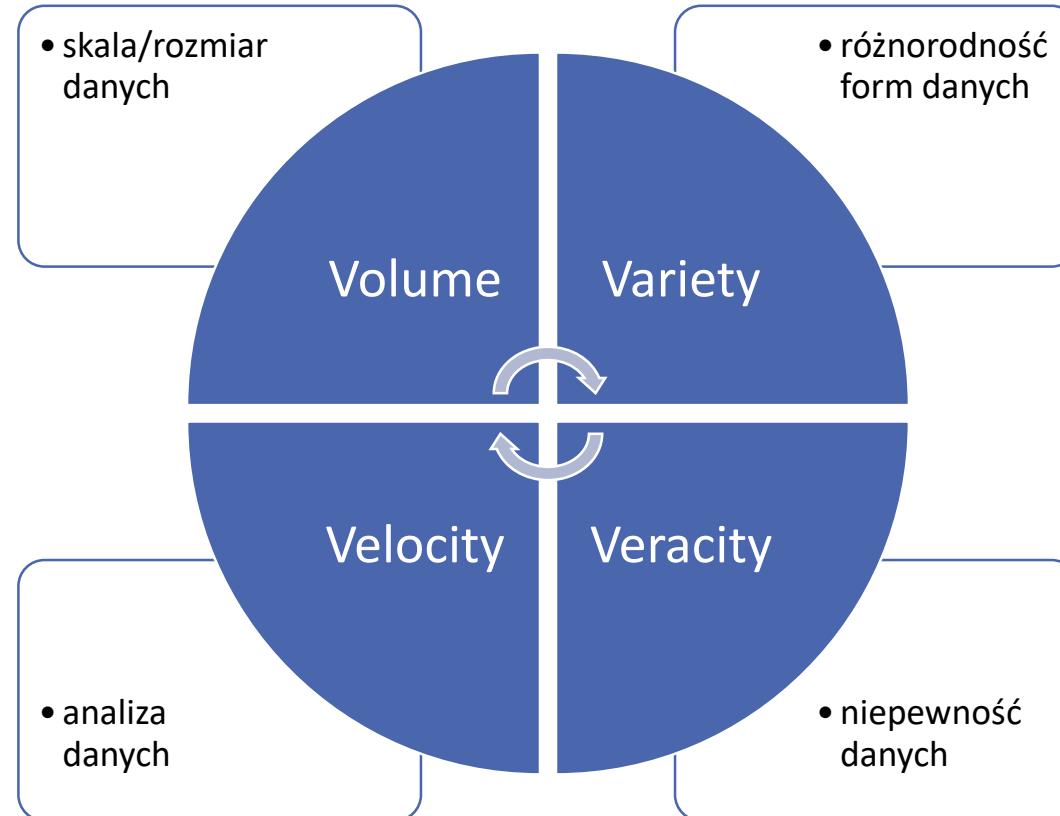
### Przykład 4: integracja federacji

- federacja danych – zbiór niezależnych, ale spójnych semantycznie hurtowni danych
- zadanie:
  - dane: zapytanie użytkownika końcowego
  - szukane: wynik zintegrowany na podstawie odpowiedzi z różnych hurtowni składowych
- problemy:
  - znalezienie odpowiadających atrybutów w różnych  $H_i$
  - integracja odpowiedzi



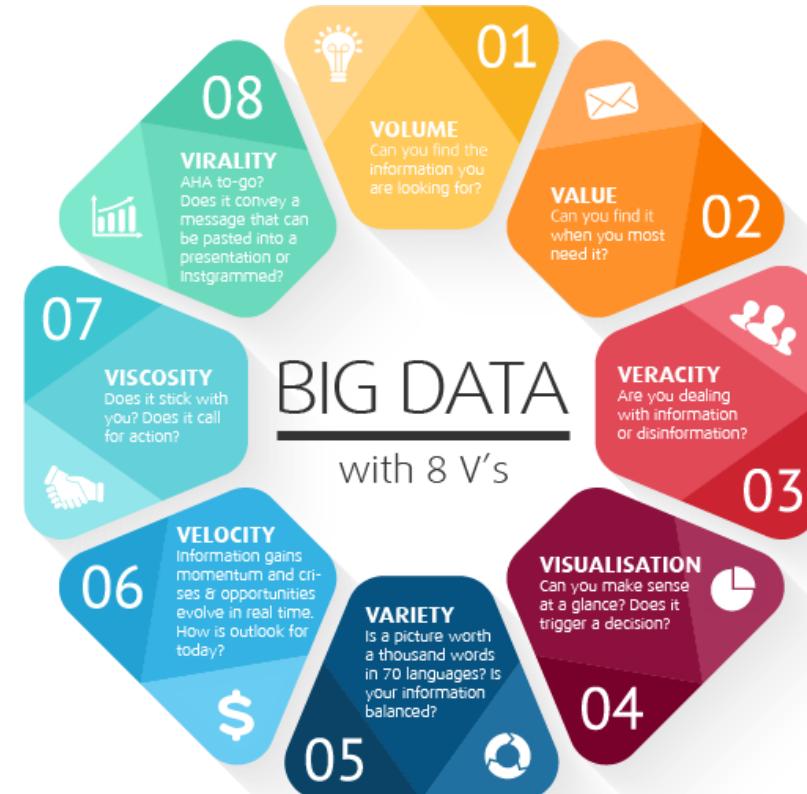
## Zastosowania hurtowni danych

### Przykład 5: Big Data



# Big Data – 8V

- Volume – wielkość danych
- Variety – mnogość formatów danych
- Velocity – szybkość napływu danych
- Veracity – prawdziwość danych
- Value – wartość biznesowa
- Visualisation – możliwość prezentacji
- Viscosity – wpływ na działania
- Virality – możliwość publikowania





*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Korzyści wdrożenia hurtowni danych

- odciążenie systemów transakcyjnych
- poprawa jakości analizowanych danych
- przechowywanie danych o długim horyzoncie czasowym
- łączenie danych pochodzących z różnych systemów transakcyjnych
- udostępnienie danych dla wszystkich potrzebujących



Fundusze  
Europejskie  
Wiedza Edukacja Rozwój



Politechnika Wrocławskiego

Unia Europejska  
Europejski Fundusz Społeczny



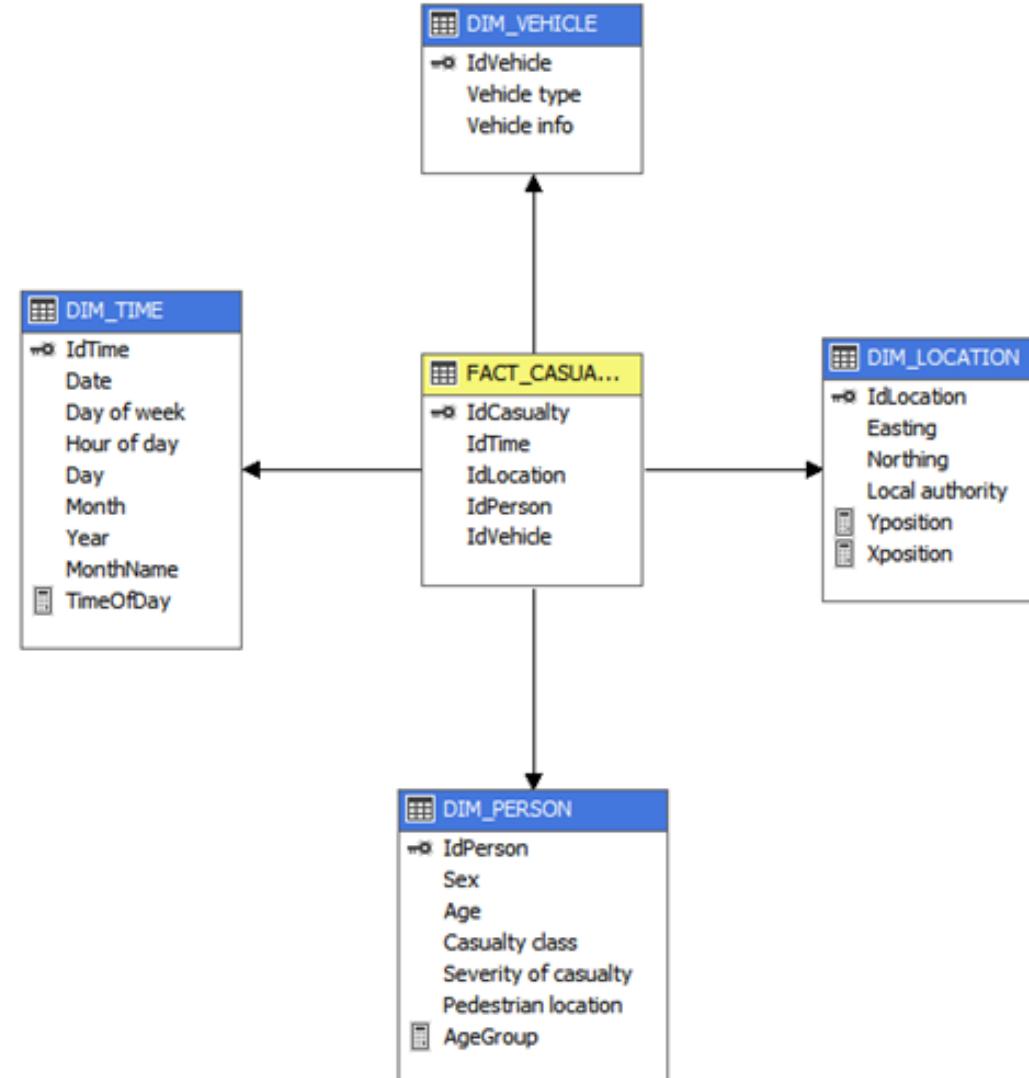
*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Projekt – uwagi

*(od prostej hurtowni do analizy biznesowej)*

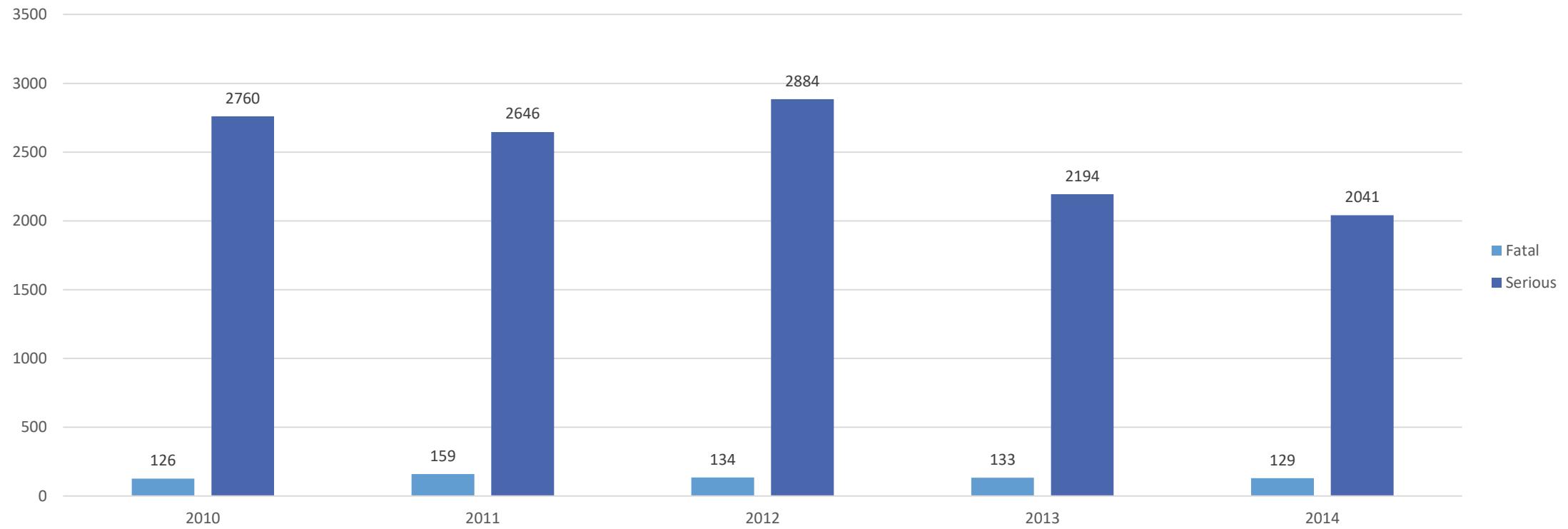


*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*



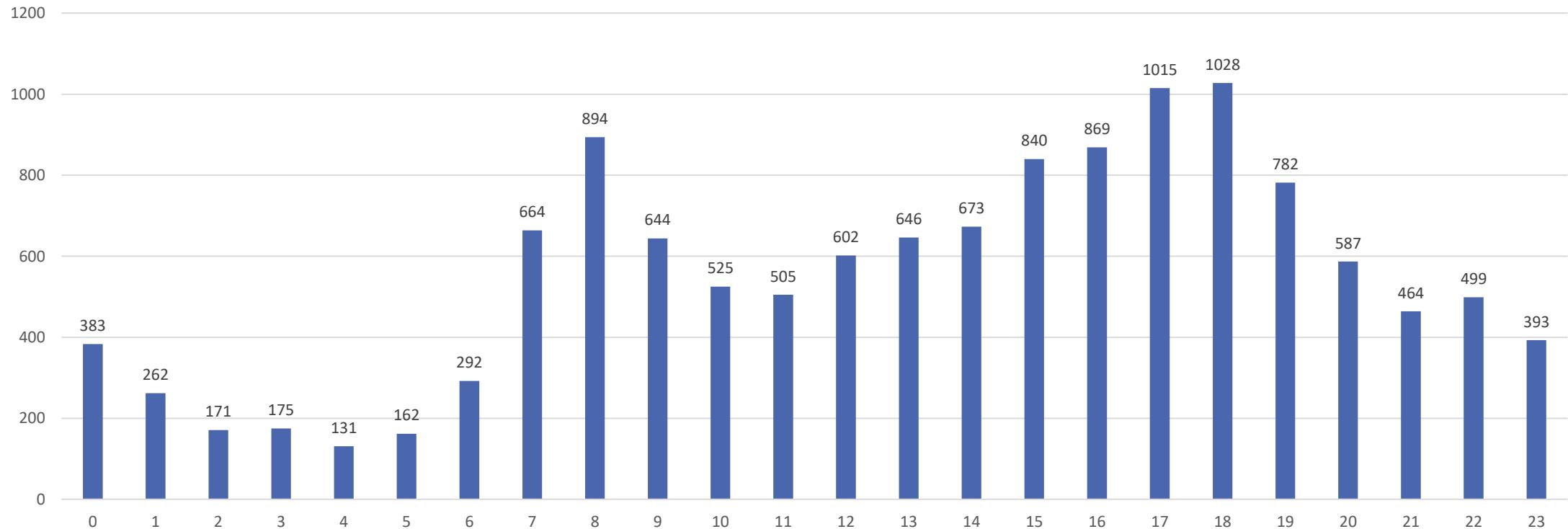
*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Liczba ofiar na przestrzeni lat



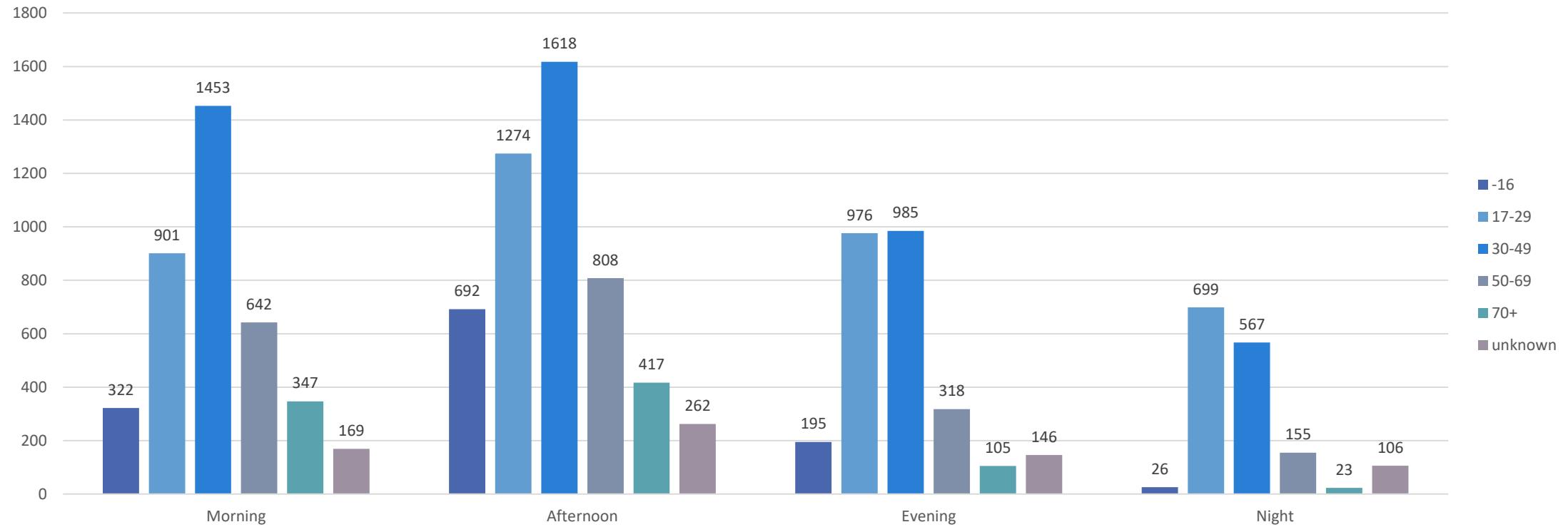
*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Liczba ofiar w zależności od godziny



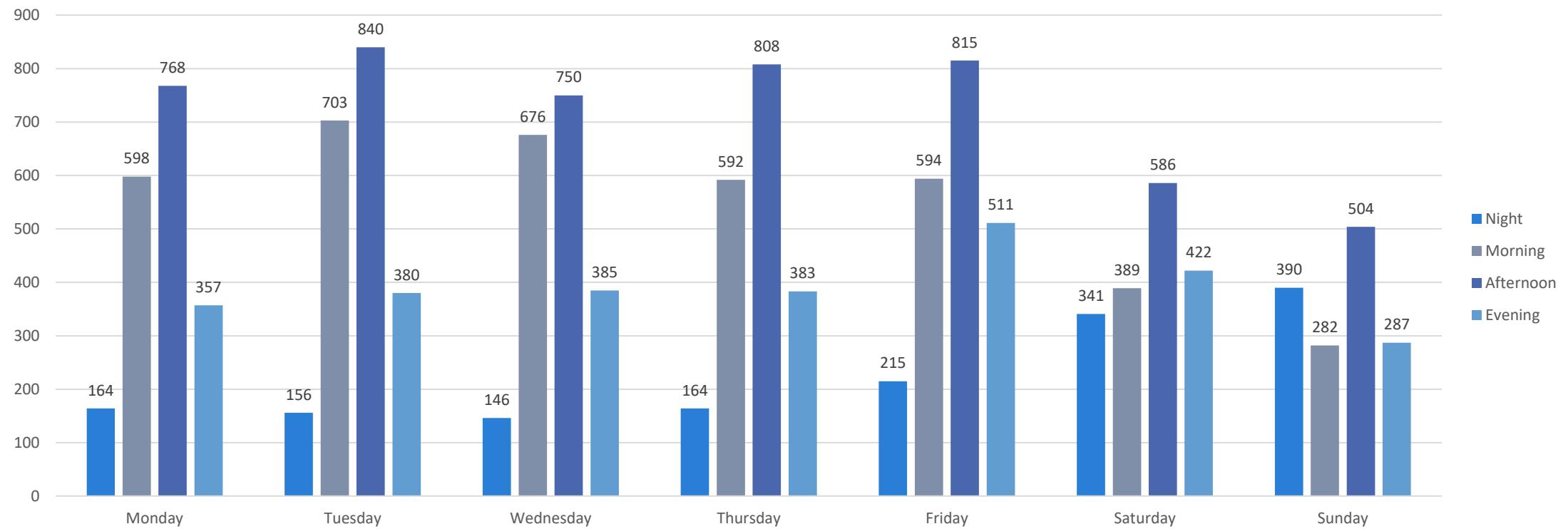
„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”

## Wiek ofiar w zależności od pory dnia



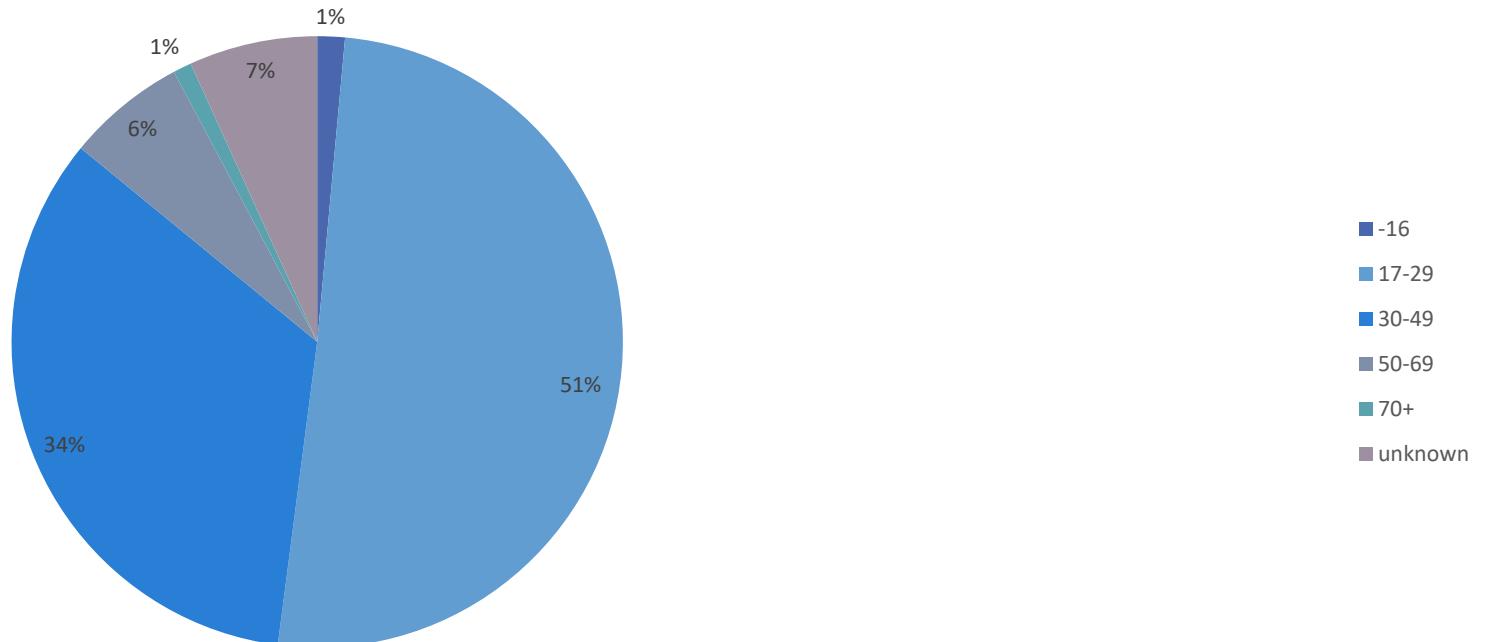
„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”

# Liczba ofiar w zależności od pory dnia w ciągu tygodnia



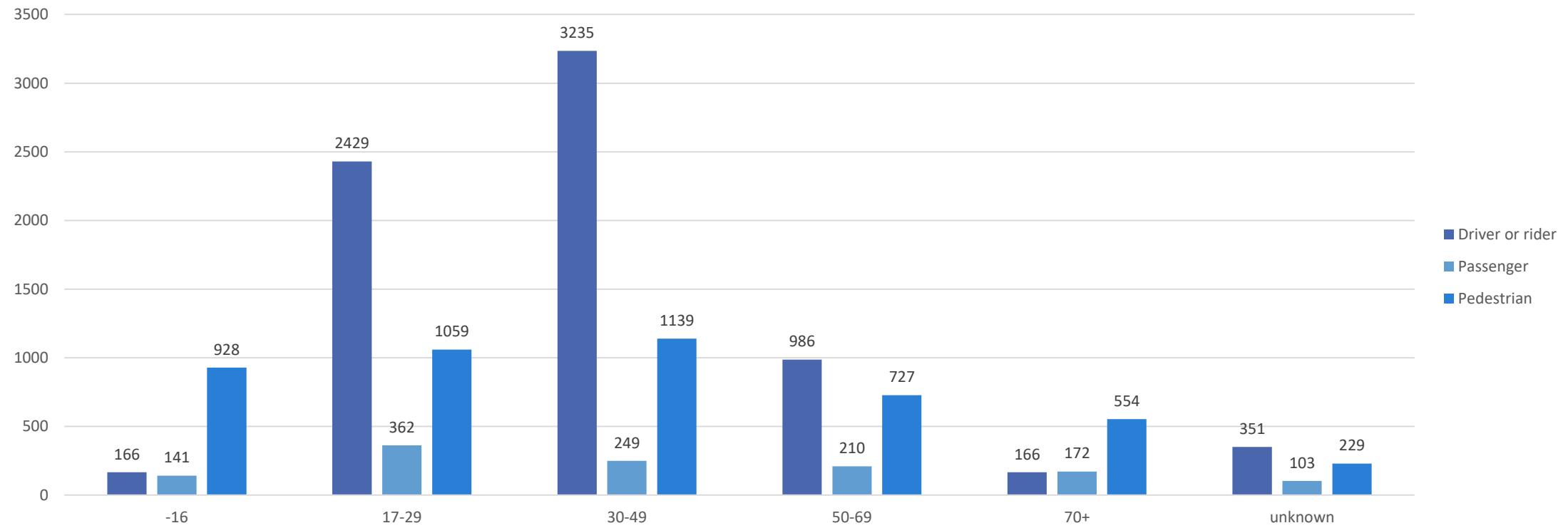
„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”

## Procent ofiar w zależności od wieku w weekendowe noce



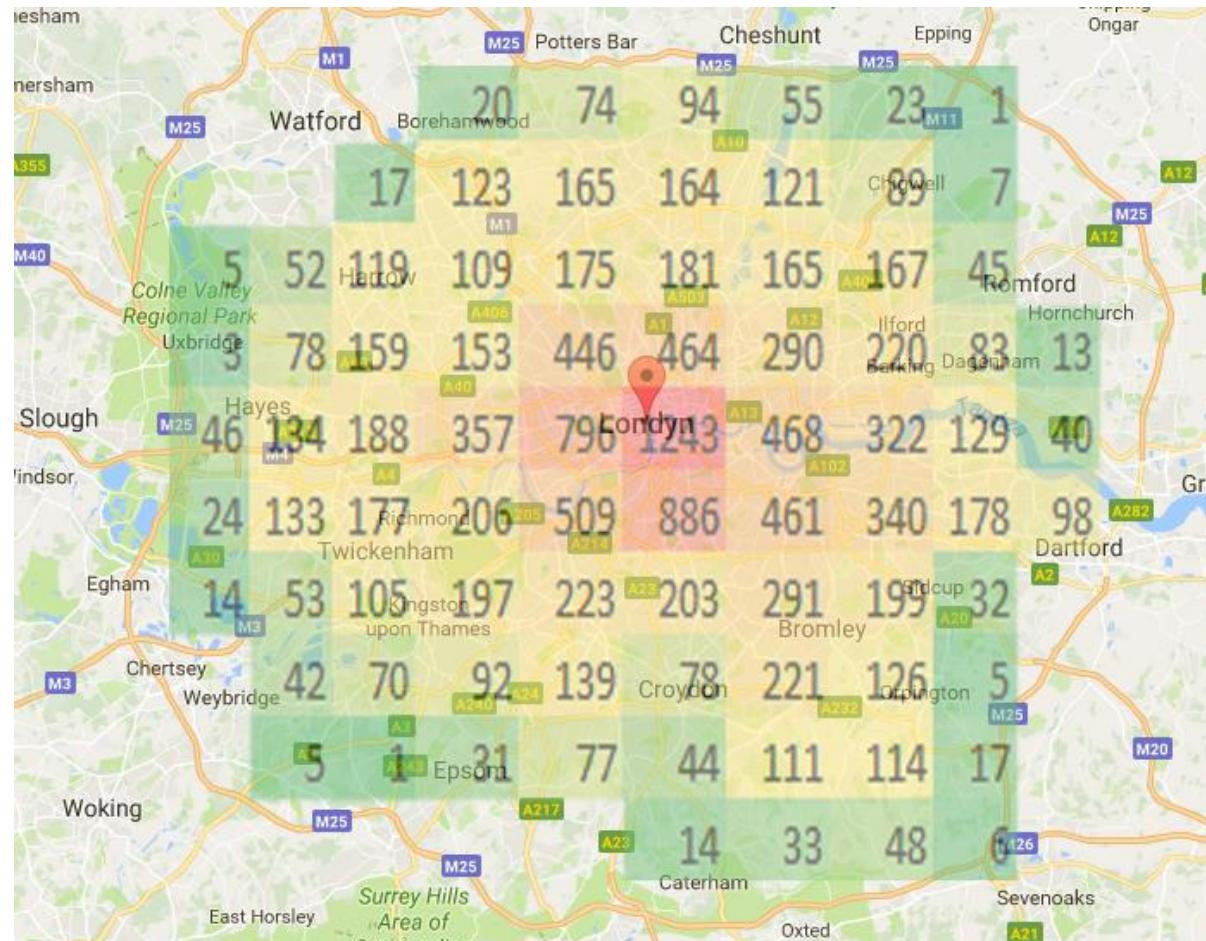
„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”

## Rodzaj ofiary w zależności od wieku



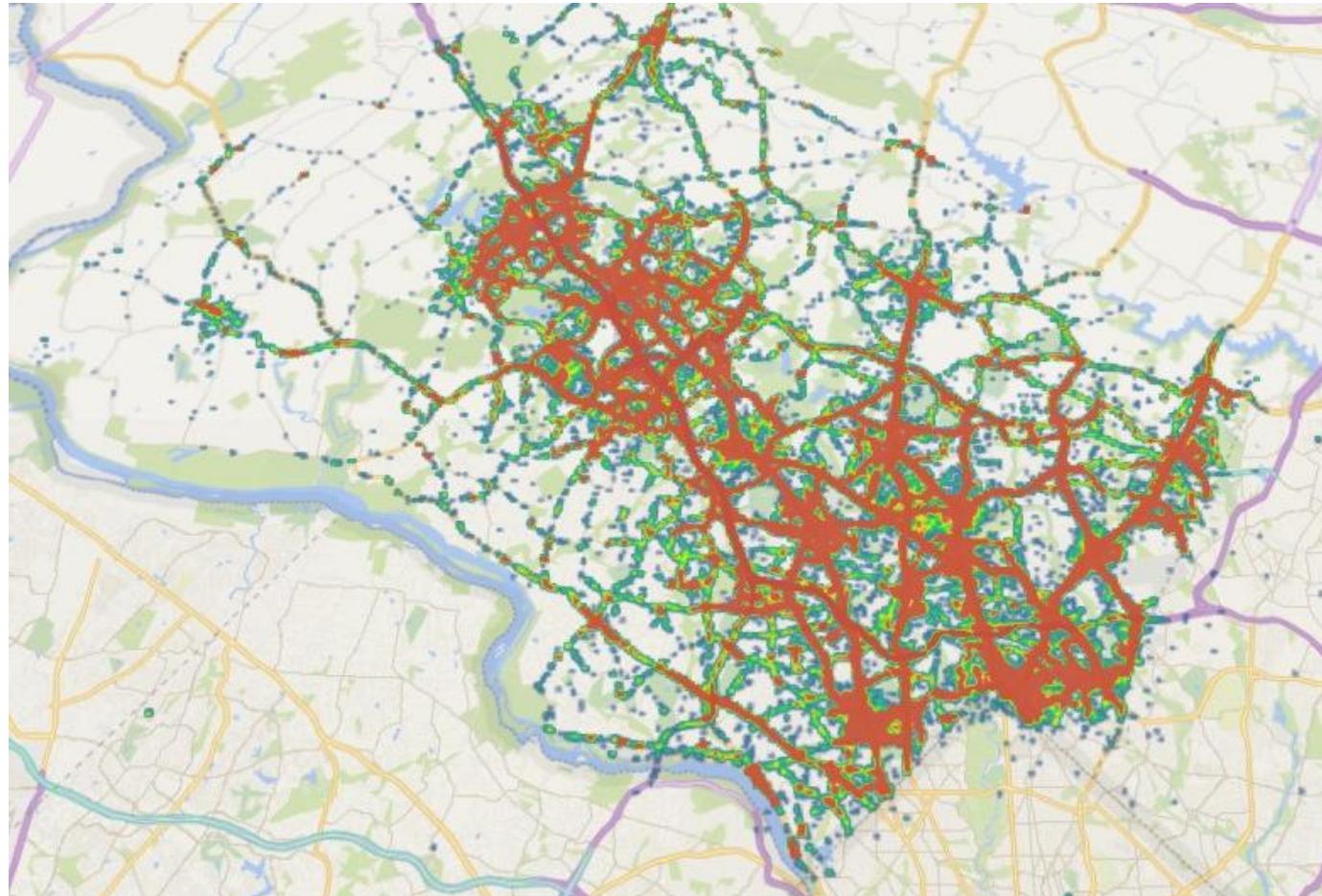
„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”

# Mapa ciepła



*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Wykroczenia drogowe - Montgomery





Fundusze  
Europejskie  
Wiedza Edukacja Rozwój



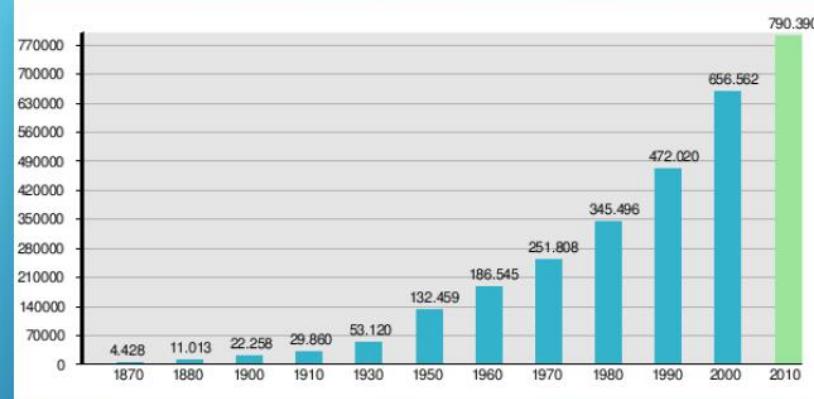
Politechnika Wrocławskiego

Unia Europejska  
Europejski Fundusz Społeczny



„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”

# Wypadki Austin



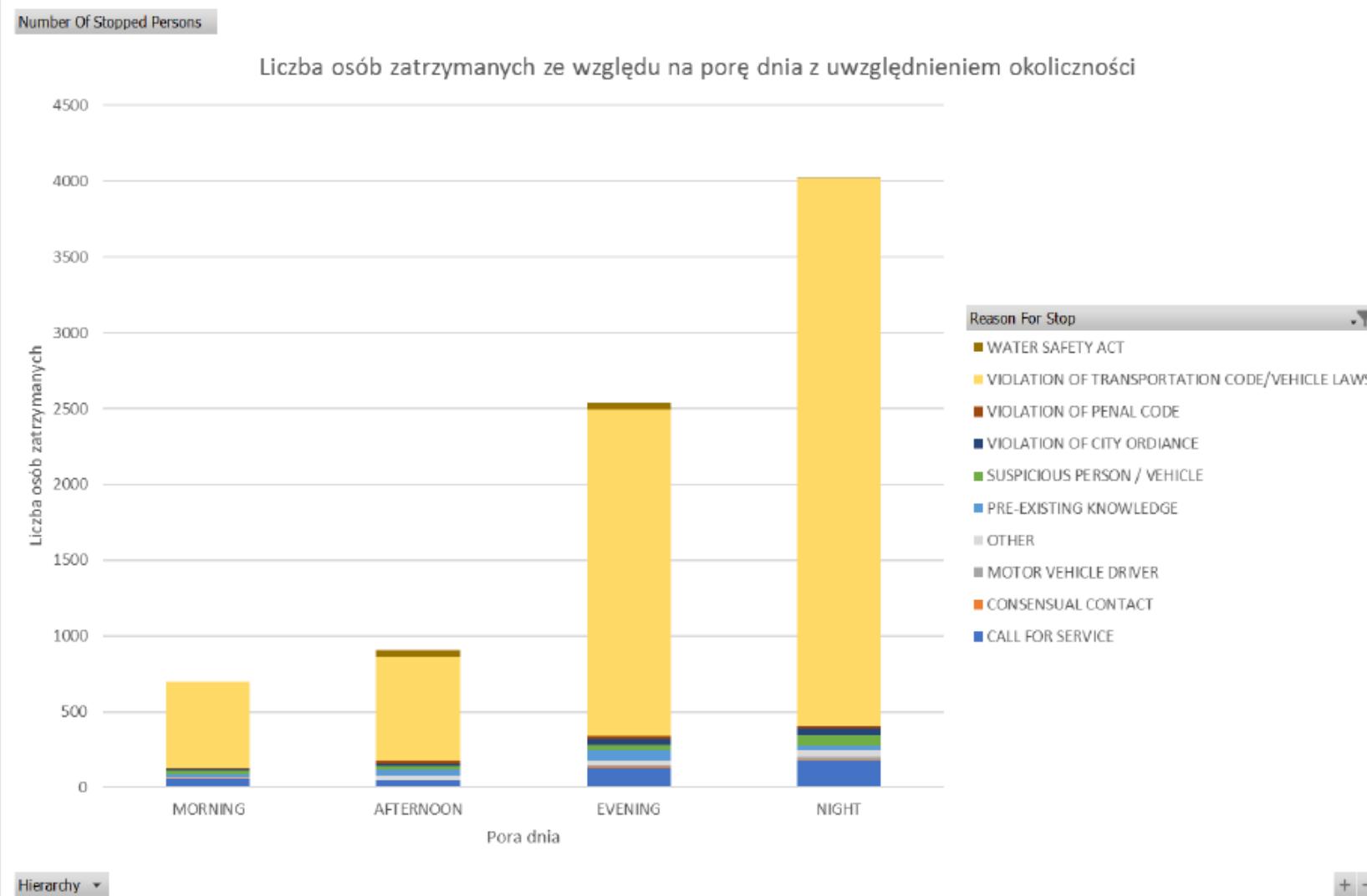
	Austin	Wrocław
liczba mieszkańców	912 791	637 683
powierzchnia [km <sup>2</sup> ]	704	292,82
gęstość ludności [os/ km <sup>2</sup> ]	1 065,04	2176



## INFORMACJE O MIEŚCIE



**„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”**





Fundusze  
Europejskie  
Wiedza Edukacja Rozwój



Politechnika Wrocławskiego

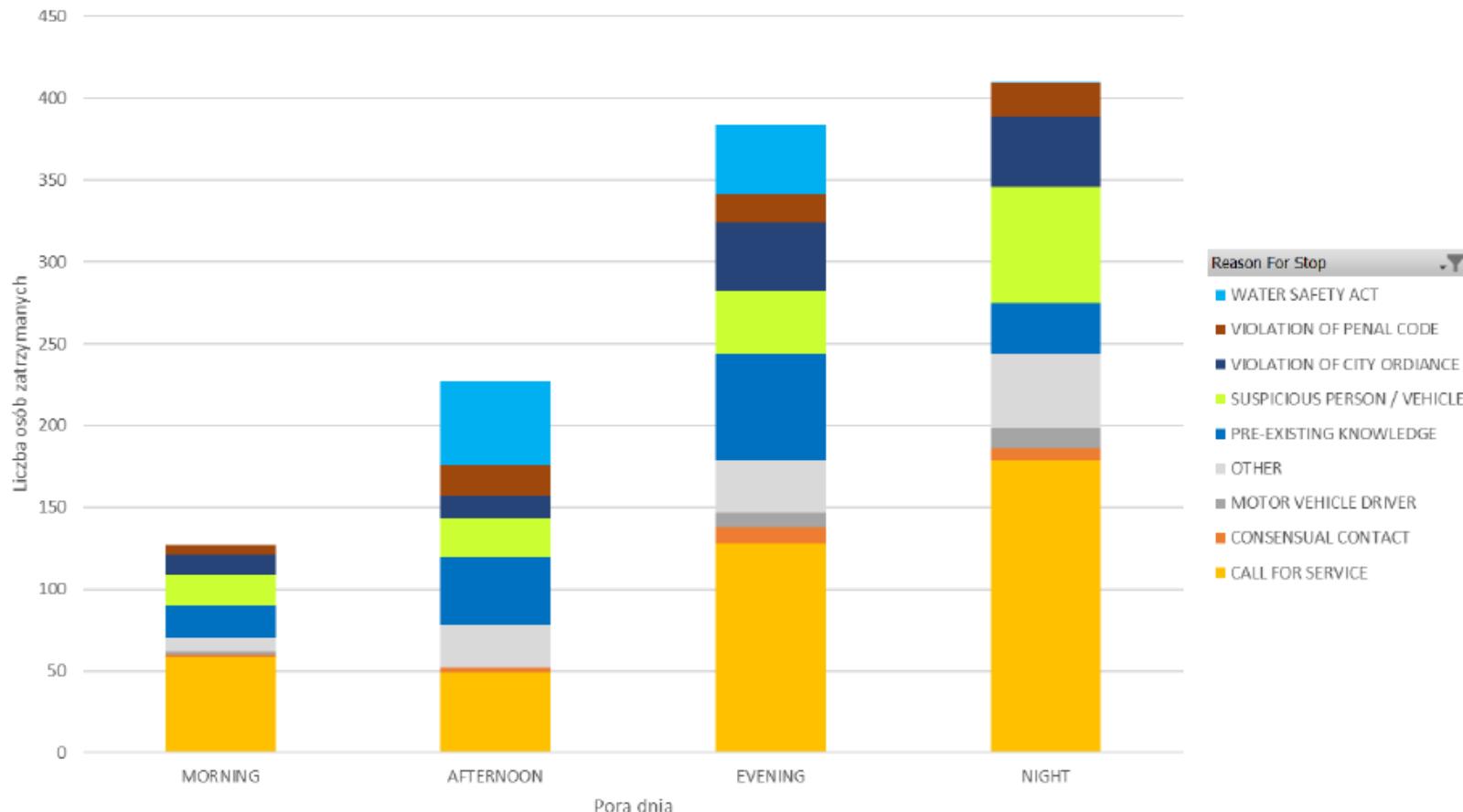


Unia Europejska  
Europejski Fundusz Społeczny

### „ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”

Number Of Stopped Persons

Liczba osób zatrzymanych ze względu na porę dnia z uwzględnieniem okoliczności



Hierarchy ▾

+

-



Fundusze  
Europejskie  
Wiedza Edukacja Rozwój



Politechnika Wrocławskiego

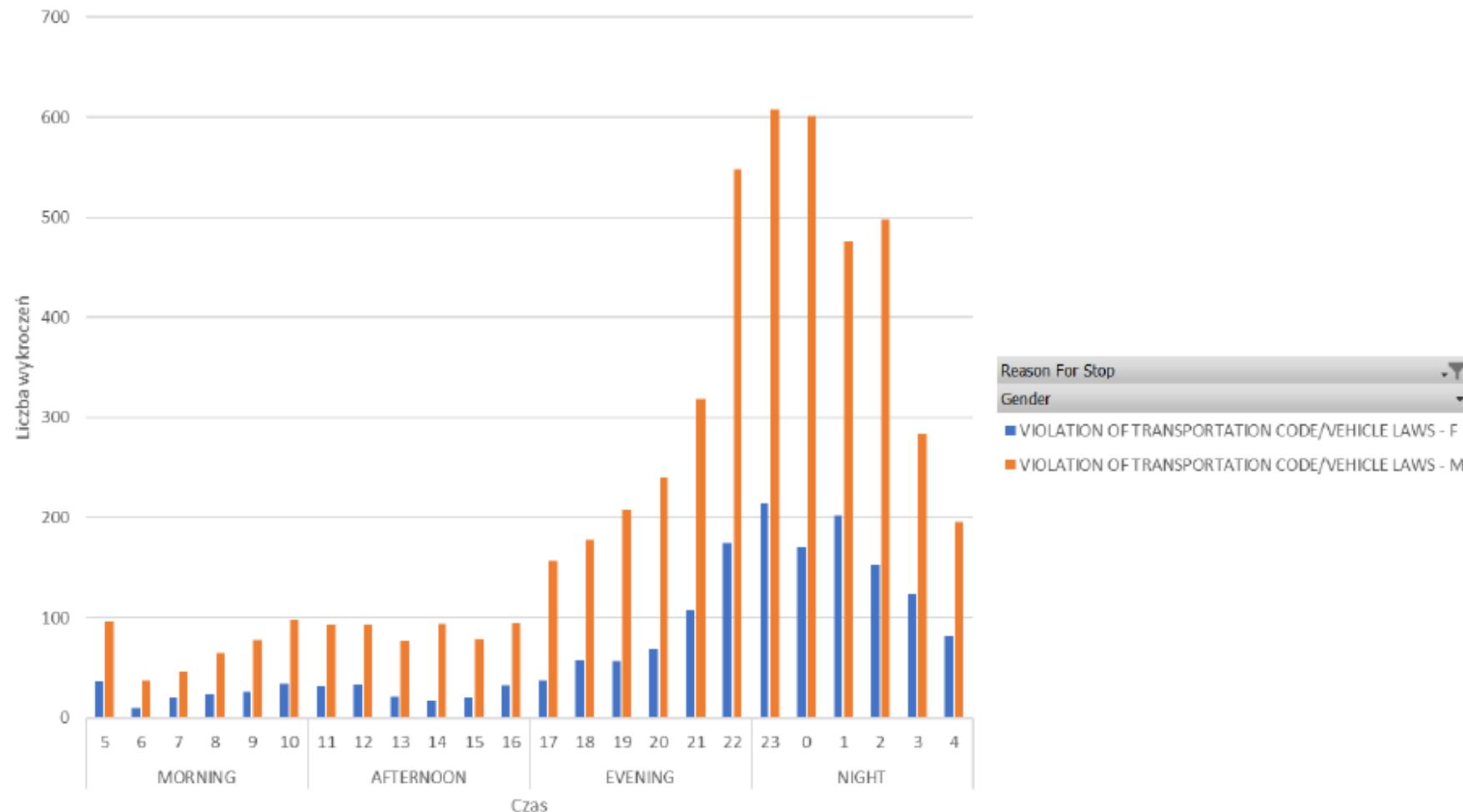
Unia Europejska  
Europejski Fundusz Społeczny



### „ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”

Number Of Stopped Persons

#### Wykroczenia w ruchu drogowym w ciągu dnia z uwzględnieniem płci

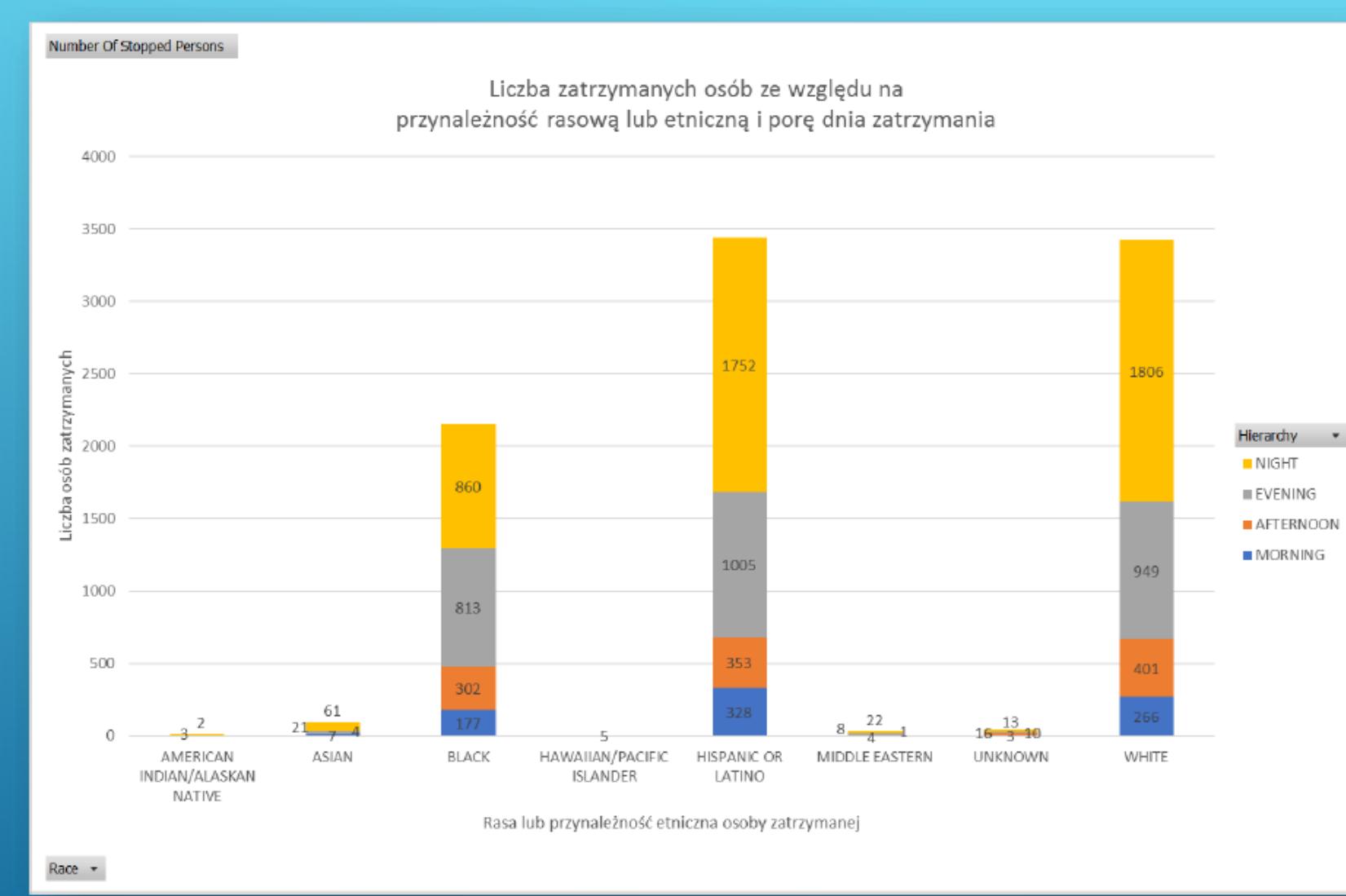


Hierarchy ▾

+ -



„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”



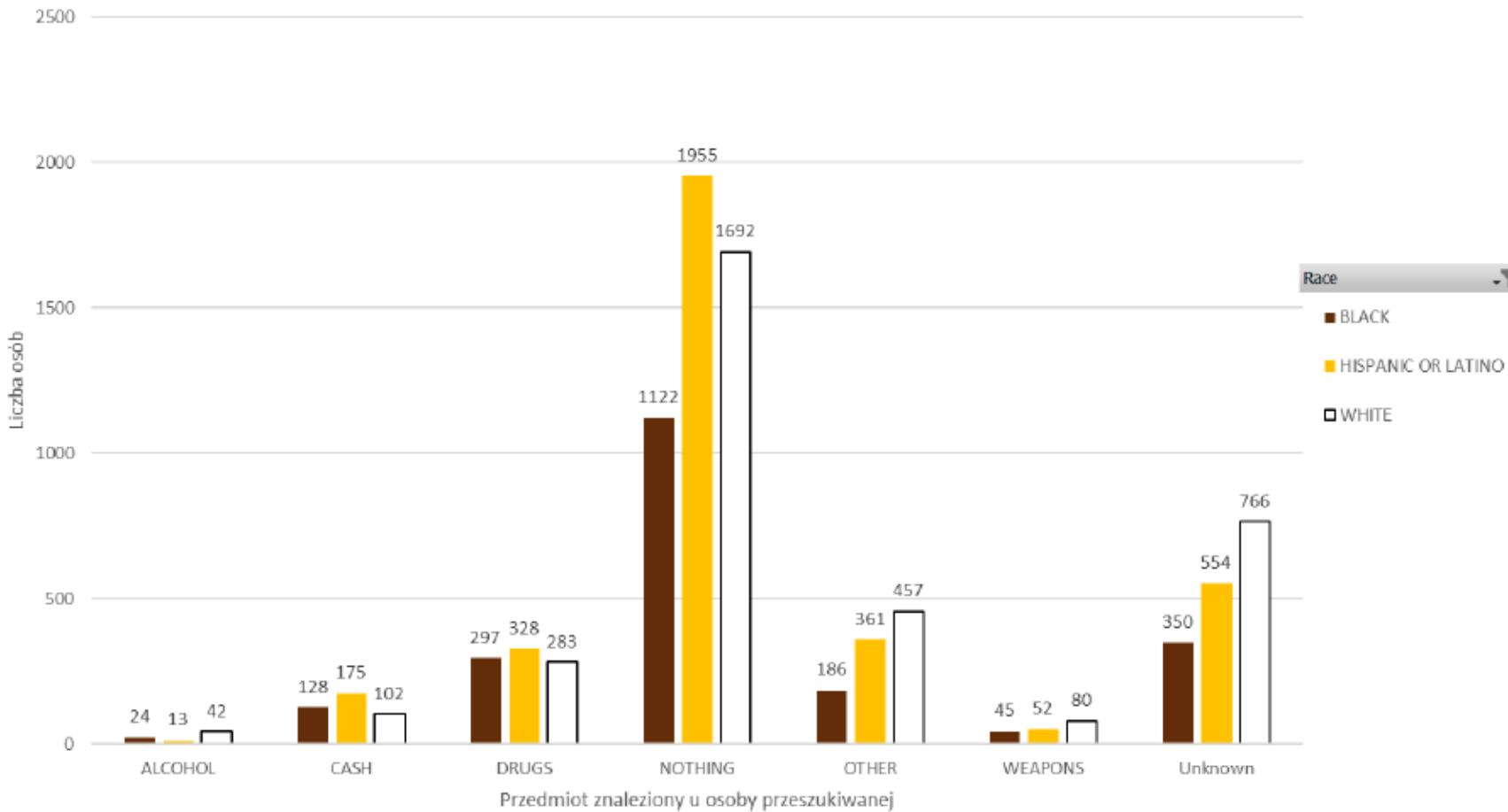
LUDNOŚĆ AUSTIN	
Biała	48,70%
Latynosi i Hiszpanie	54,70%
Afroamerykanie	8,10%
Azjaci	6,30%
Indianie i rdzeni mieszkańcy Alaski	0,90%
Hawajczycy i wyspiarze Pacyfiku	0,10%



**„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”**

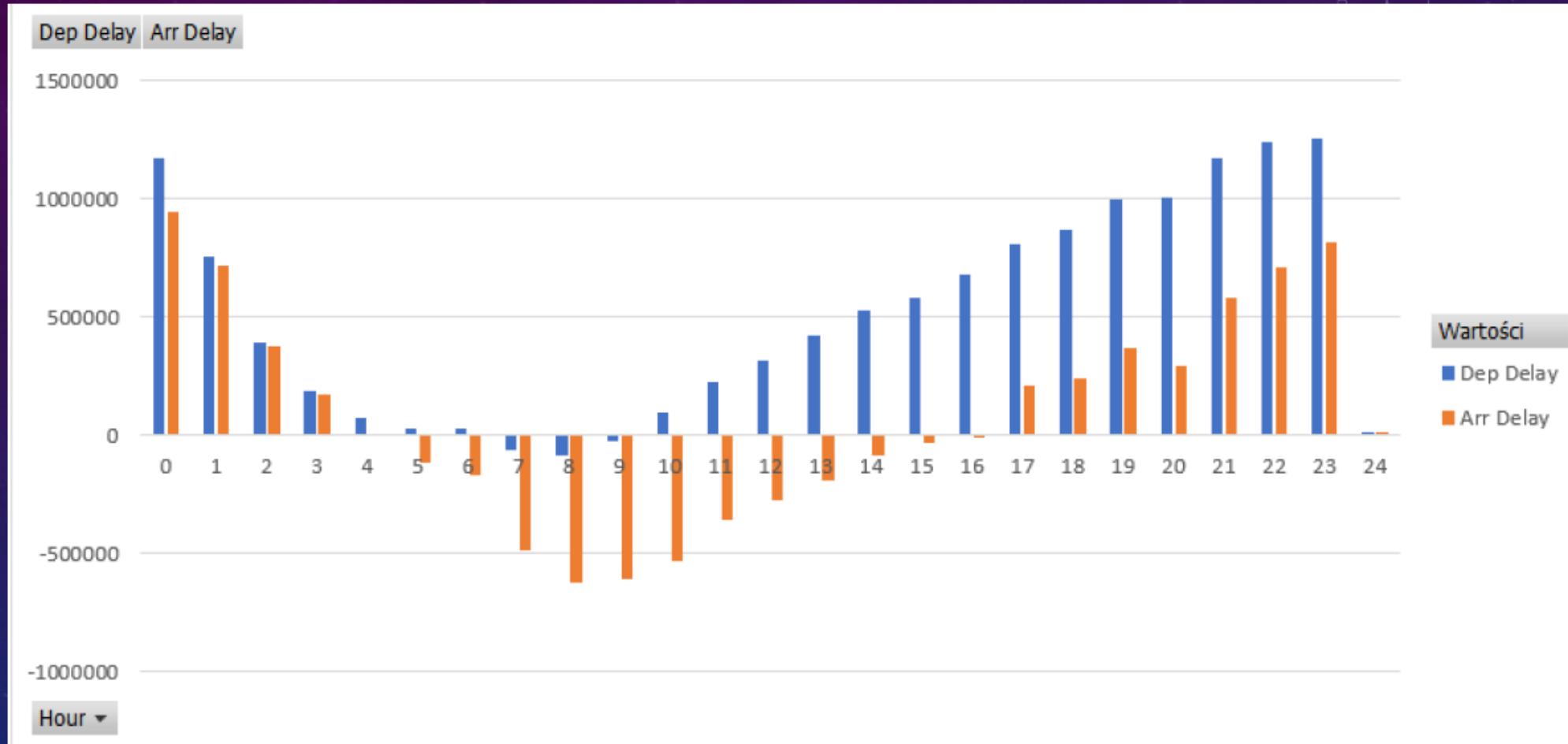
Number Of Stopped Persons

Liczba osób przeszukanych, u których znaleziono określone przedmioty z uwzględnieniem  
przynależności etnicznej osoby zatrzymanej



Item Found ▾

# Rozkład opóźnień odlotów i przylotów w zależności od godziny

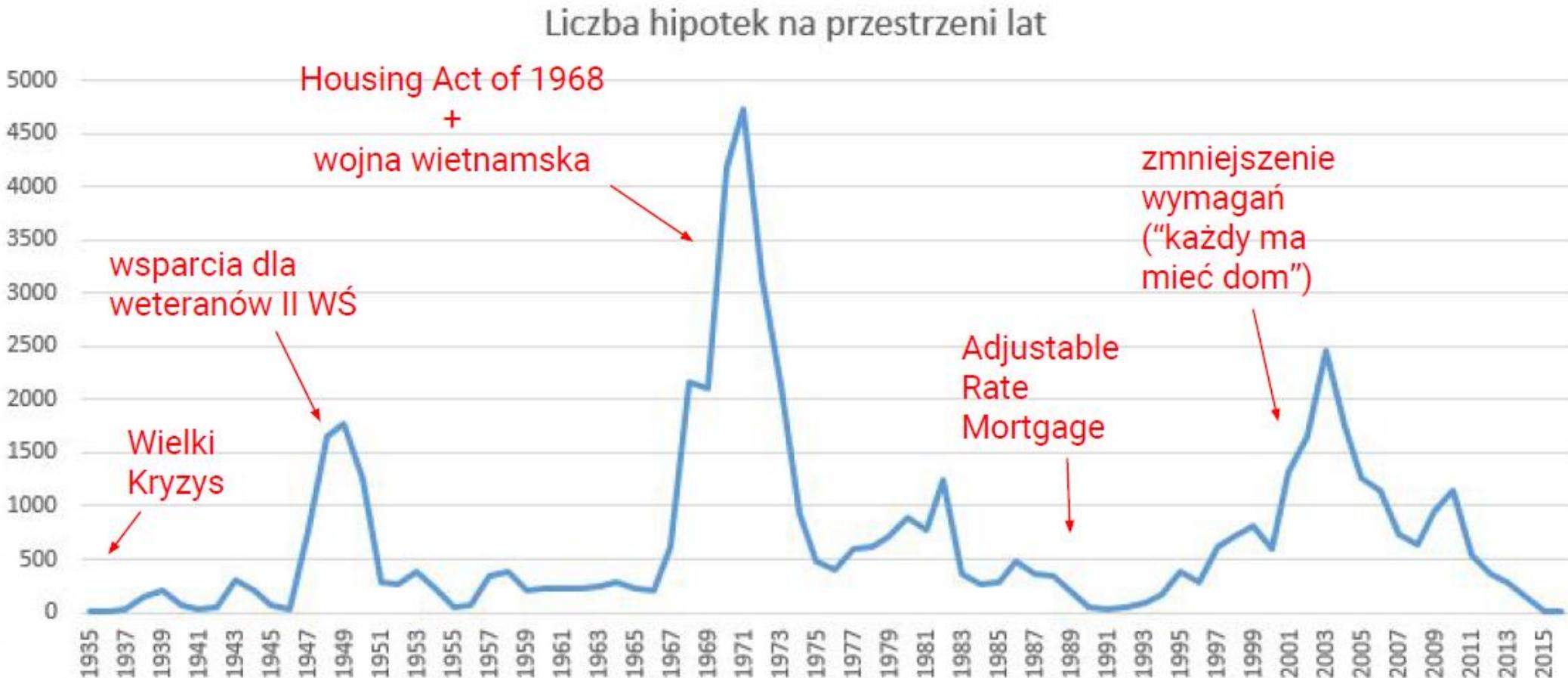


Dep Delay – suma opóźnień odlotów.

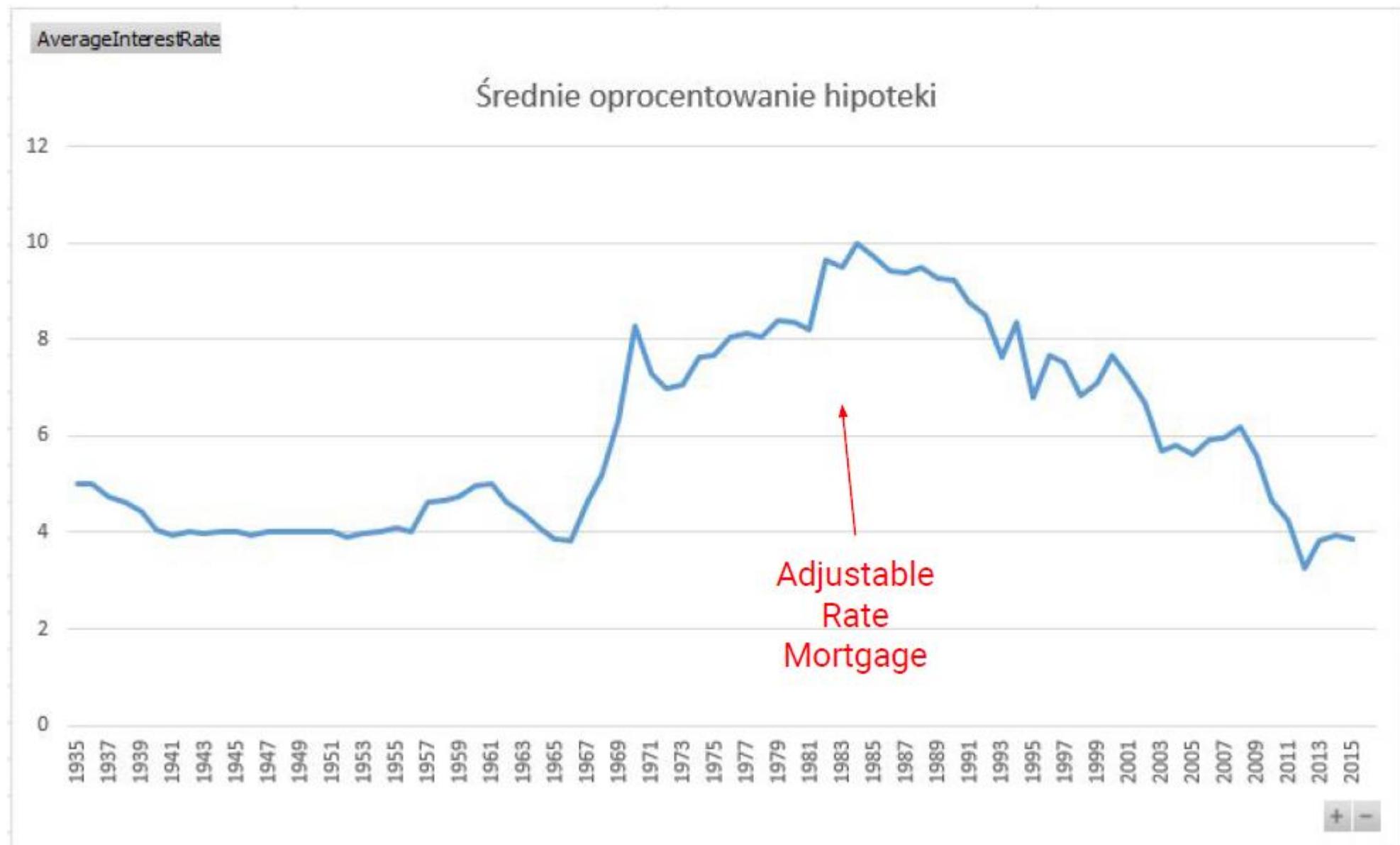
Arr Delay – suma opóźnień przylotów.



FACT Mortgages Count



Product	Rate	Change	Last week
30 year fixed refi	3.80%	▲ 0.06	3.74%

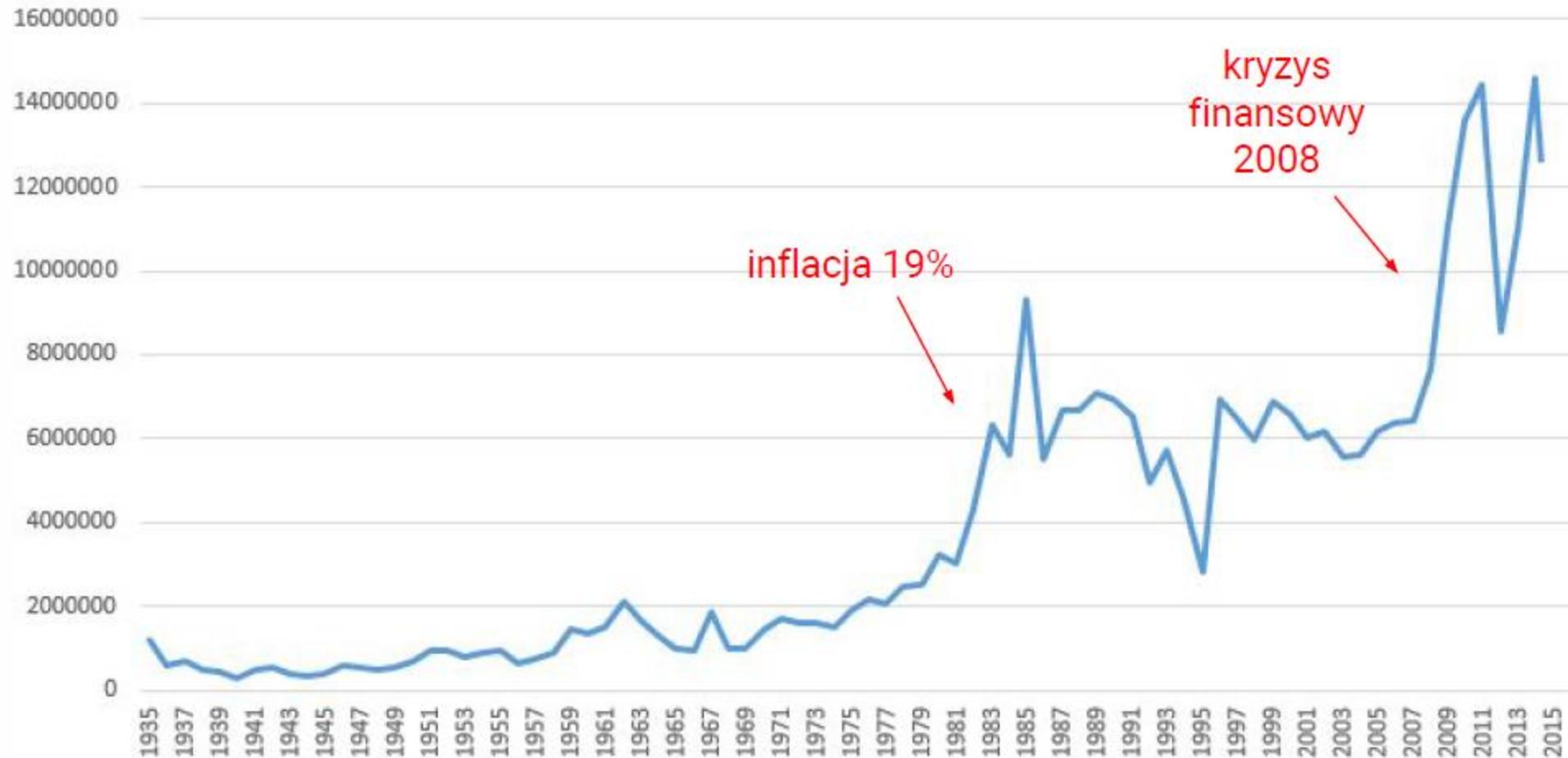




„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”

AverageOriginalMortgageAmount

Średnia początkowa kwota pożyczki





Fundusze  
Europejskie  
Wiedza Edukacja Rozwój



Politechnika Wrocławska

Unia Europejska  
Europejski Fundusz Społeczny



*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Hurtownie danych

Dziękuję za uwagę

dr inż. Marcin Maleszka



Fundusze  
Europejskie  
Wiedza Edukacja Rozwój



Politechnika Wrocławska

Unia Europejska  
Europejski Fundusz Społeczny



*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Hurtownie danych

**Webowe panele zarządzania – dashboards**

**dr inż. Marcin Maleszka**



# Dashboard

- Pulpit nawigacyjny:
  - rodzaj graficznego interfejsu użytkownika
  - zapewnia szybki przegląd kluczowych wskaźników wydajności związanych z określonym celem lub procesem biznesowym.



# Dashboard – pulpit nawigacyjny

- Aplikacja do analizy danych biznesowych
- Konsolidacja i prezentacja na jednym ekranie wiele elementów
- Zastosowania:
  - raportowanie
  - analiza
  - monitorowanie
  - kontrola
- Prezentacja kluczowych wskaźników wydajności (KPI) w postaci:
  - raportów statycznych lub dynamicznych, tabel, wykresów
- Podkreślenie nieprawidłowości

„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”

# Wyzwania panelu nawigacyjnego

- Panel nawigacyjny powinien:
  - odpowiadać **potrzebom biznesowym** przedsiębiorstwa
  - wyświetlać wszystkie wymagane informacje na jednym ekranie **w czytelny sposób**
  - prezentować informacje w sposób **prosty, zwięzły, jasny** i wyjątkowo dobrze zorganizowany
  - podkreślać **podsumowania i wyjątki**
  - być **wydajny** – wykorzystywać tylko te dane, które są chwilowo potrzebne
  - być dostosowany do **uprawnień i potrzeb** różnych typów użytkowników
  - prezentować **różne widoki danych**
  - zawierać różnorodność właściwie dobranych **efektów wizualnych**

# Cechy dashboardów

- Panel powinien zawierać kluczowe wskaźniki wydajności:
  - ważne z punktu widzenia biznesu: strategii i celów wydajnościowych
  - „rozwijalne” – te, dla których można sprawdzić szczegóły i przeprowadzić dalszą analizę
  - porównywalne ze standardami biznesowymi
  - odpowiadające za opłacalność przedsiębiorstwa
  - możliwe do porównania ze średnimi, medianami, wartościami oczekiwanyimi
- Użytkownik jest w stanie jednocześnie skupić się na 3-7 wskaźnikach
- Całkowita liczba KPI na jednym ekranie nie powinna być większa niż 20

# Rodzaje dashboardów

- Dashboard menedżerski
  - panel zawierający podstawowe raporty
  - popularna „sygnalizacja” stanu firmy
  - stan przedsiębiorstwa na tle zdefiniowanych celów czy innych punktów odniesienia (konkurencja, porównanie z poprzednimi okresami, inne rynki, itp.)
- Dashboard analityczny
  - wizualizacja danych
  - możliwość sprawdzenia przyczyn niepokojącego wyniku
  - analiza w głąb

# Dashboard menedżerski a analityczny

	<b>Dashboard menedżerski</b>	<b>Dashboard analityczny</b>
Cel	Aktualne wyniki Wyróżnienie wyjątków	Możliwość szukania przyczyn u źródła
Użytkownik	Menedżer	Menedżer lub analityk
Interakcje analityczne	Podstawowa, głównie zdefiniowane opcje podglądu w głąb	Szeroki zakres możliwości
Częstotliwość aktualizacji	Raporty cykliczne	Raporty ad-hoc
Forma wizualna	Wykresy i tabele	Główne wykresy

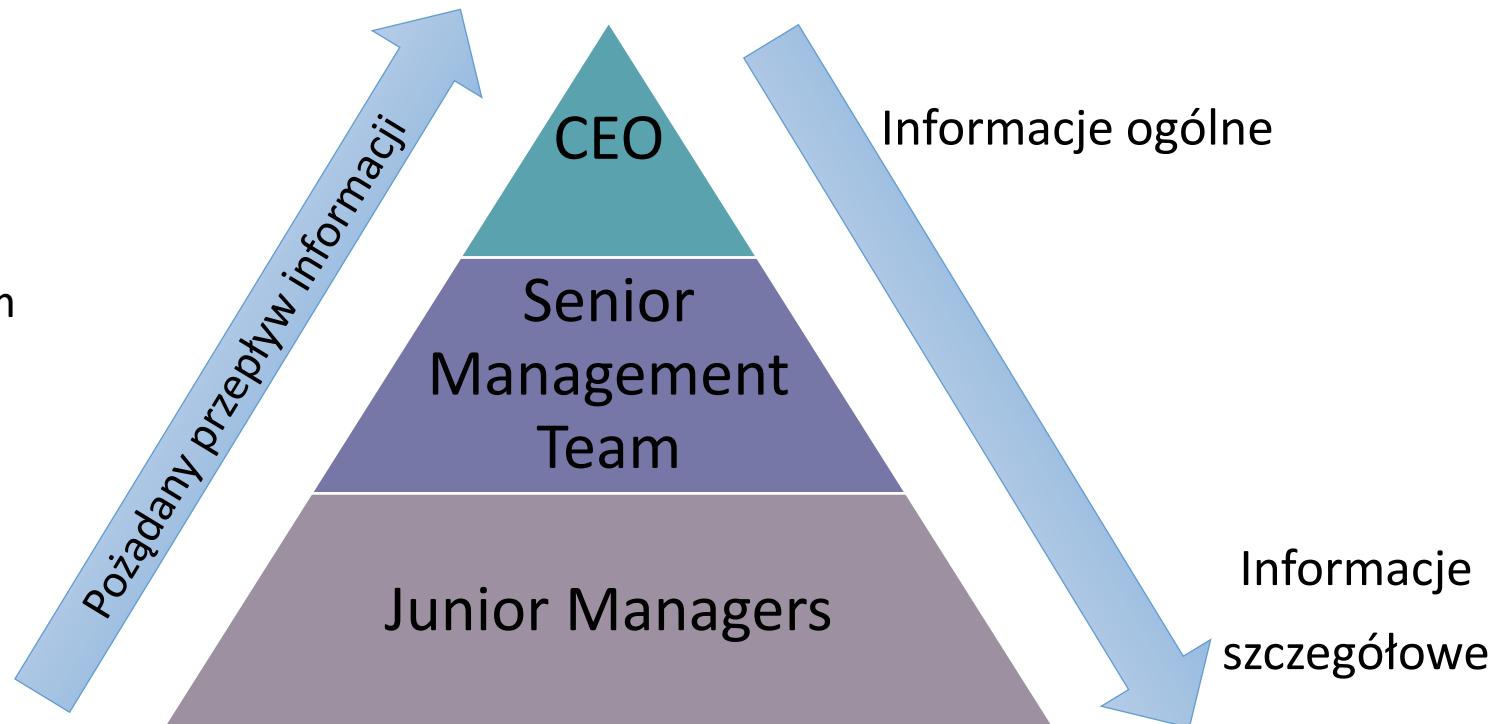
*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

## Ważne aspekty – dobre praktyki

- Dobrze zdefiniowane KPI -> lepsze decyzje biznesowe
- Projekt panelu
- Wizualizacja wskaźników
- Dobór właściwych wykresów
- Wykresy zawierające kategorie i czas
- Wykresy zawierające szczegółowe wartości
- Wykresy trendów, wykrywanie trendów
- Wykresy porównawcze, zawierające rozkład cech
- Filtrowanie i tabele przestawne

# Dobrze zdefiniowane KPI -> lepsze decyzje biznesowe

- Prezentacja właściwych KPI właściwym użytkownikom
  - które KPI prezentować pracownikowi na określonym szczeblu?
  - łatwość interpretacji
  - przydatność informacji zamiast zbioru statystyk



# Dobrze zdefiniowane KPI -> lepsze decyzje biznesowe

- Właściwości wskaźników KPI
  - KPI są zwykle okresowe, np. wartość dzienna, tygodniowa, miesięczna, itp.
  - średnia wartość wskaźnika nie jest wystarczająca
  - pożądane jest porównanie wartości tego samego wskaźnika do wartości poprzednich, założonych celów, konkurencji, itp.
  - monitorowanie zmian
- Rozróżnienie KPI od danych zarządzania:
  - KPI – cykliczne wartości, np. liczba zgłoszeń problemu określonego typu
  - dane zarządzania – wartość bieżąca, np. kolejne zgłoszenia błędów systemu

# Dobrze zdefiniowane KPI -> lepsze decyzje biznesowe

- Rodzaje KPI:
  - **ilościowe** – ściśle zdefiniowane, znane wartości krytyczne
  - **kierunkowe** – znany pożądany trend; sprawdzanie codziennej wartości nie jest tak kluczowe jak analiza w dłuższej perspektywie czasowej
  - **zapobiegawcze** (ang. actionable) – zdefiniowana wartość pożądana, którą należy osiągnąć lub jej nie przekroczyć – możliwa reakcja przy zbliżaniu się do określonego progu
  - **wskaźniki rozkładu lub kategorii** – znaczące różnice pomiędzy wartościami KPI dla różnych kategorii może świadczyć o potencjalnej nieprawidłowości

# Projektowanie panelu nawigacyjnego

- Panel ma być:
  - czytelny,
  - przejrzysty,
  - atrakcyjny wizualnie,
  - itp.,
- ALE: najważniejsza jest jego **zawartość**
- Widok kluczowych wskaźników wydajności organizacji ma być łatwy do zrozumienia
- Ma pozwalać odbiorcy szybko dostrzegać możliwości/zagrożenia i podejmować decyzje

# Projektowanie panelu nawigacyjnego

- Cele szczegółowe:
  - porównania
  - prezentacja trendu
  - prezentacja rozkładów
  - prezentacja zależności pomiędzy etapami procesu a końcowym efektem
- Ograniczenie poziomu szczegółowości -> poprawa czytelności
  - różne wykresy (2-3 typy) prezentujące te same dane
  - różne wykresy dla kolejnych serii danych (zamiast wszystkich serii na jednym wykresie)
  - „przybliżenia” danych – np. wykres trendu w dłuższej perspektywie i rozwinięcie ostatnich kilku wartości

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Wizualizacja wskaźników KPI

- Jakie wskaźniki ma zawierać panel?
  - jakie miary wybrać? które są ważne?
  - co chcę pokazać: informację pozytywną czy negatywną?
  - czy są zdefiniowane progi, wartości pożądane, graniczne, itp.?
  - jaki jest charakter wskaźnika: chcę analizować wartość, trend, rozkład, itp.?
  - z czym chcę porównać wybrany wskaźnik?
- Które wskaźniki są najważniejsze, a które tylko uzupełniają obraz stanu organizacji?

*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Dobór właściwych wykresów

- Jaki typ wykresu będzie najlepszy?
  - wykres czasowy (liniowy) z linią trendu
  - porównanie rozkładu kategorii (kołowy, kolumnowy)
  - tarcze (ang. gauges and dials)
  - karty wyników
  - tabele postępu
  - surowe dane
  - wykresy porównawcze
  - itp.
- Wybór sposobu prezentacji zależy od zamierzonego przekazu
- Należy pamiętać o legendzie i opisach osi/kategorii

# Wykresy zawierające szczegółowe wartości

- Złota zasada:

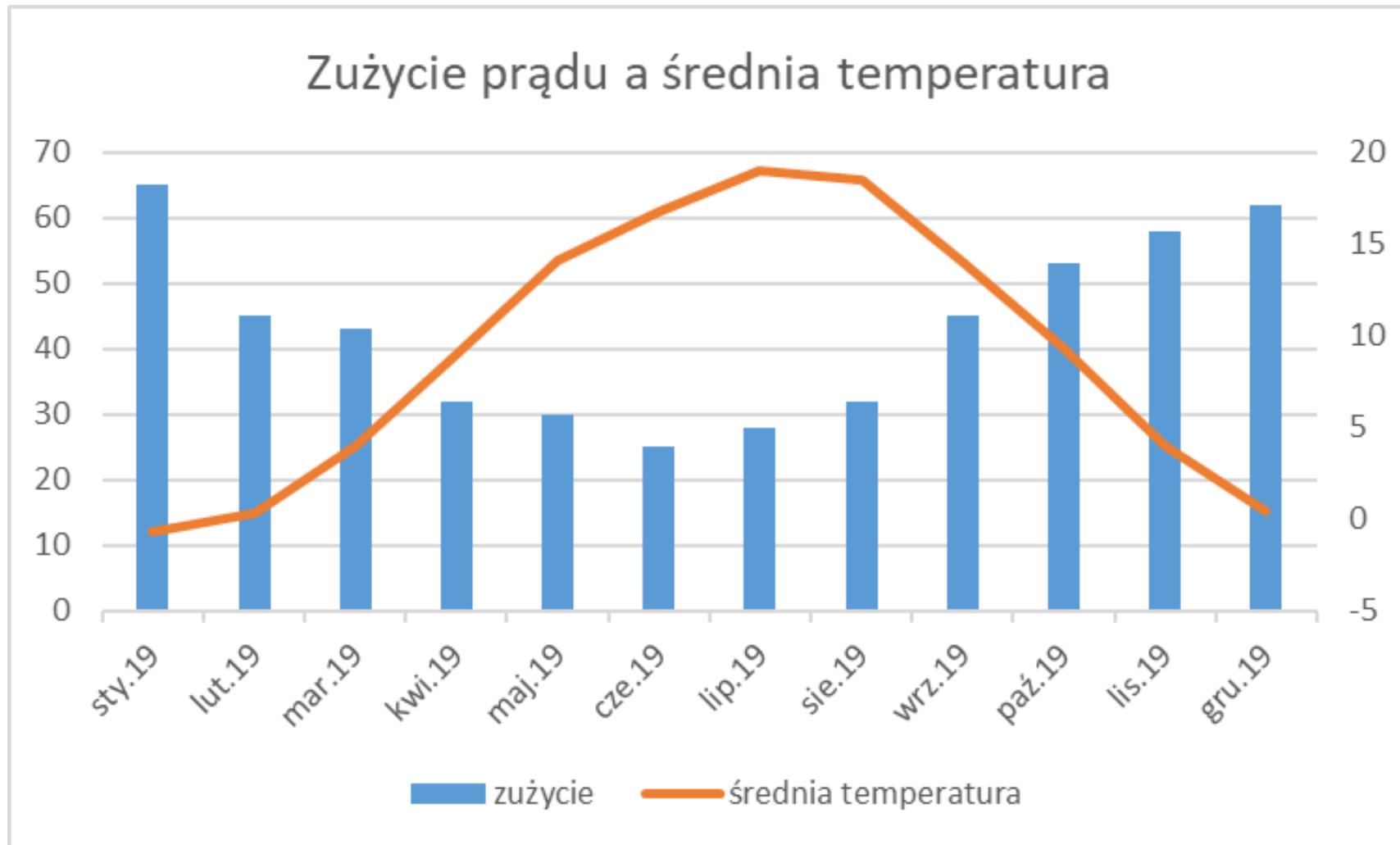
**Nie umieszczaj zbyt wielu danych na jednym wykresie**

- Dodanie etykiet danych (wartości)
  - jedna lub dwie serie danych
  - unikaj wykresów 3D z etykietami danych
  - przygotuj dwa wykresy dla porównania dłuższej i krótszej perspektywy czasowej
  - prosty wykres jest bardziej czytelny

# Wykresy trendów, wykrywanie trendów

- Dlaczego trend jest ważny?
- Zazwyczaj prezentacja trendu w postaci wykresu liniowego
  - spora część wykresu może być pusta, więc możliwa zmiana skali na osi Y
  - może być wykres warstwowy
- Łączenie wykresów, np. kolumnowego i liniowego
  - tylko dla tego samego zakresu wartości osi X
  - jasna legenda
  - odpowiedni opis skali dla każdej serii

„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”





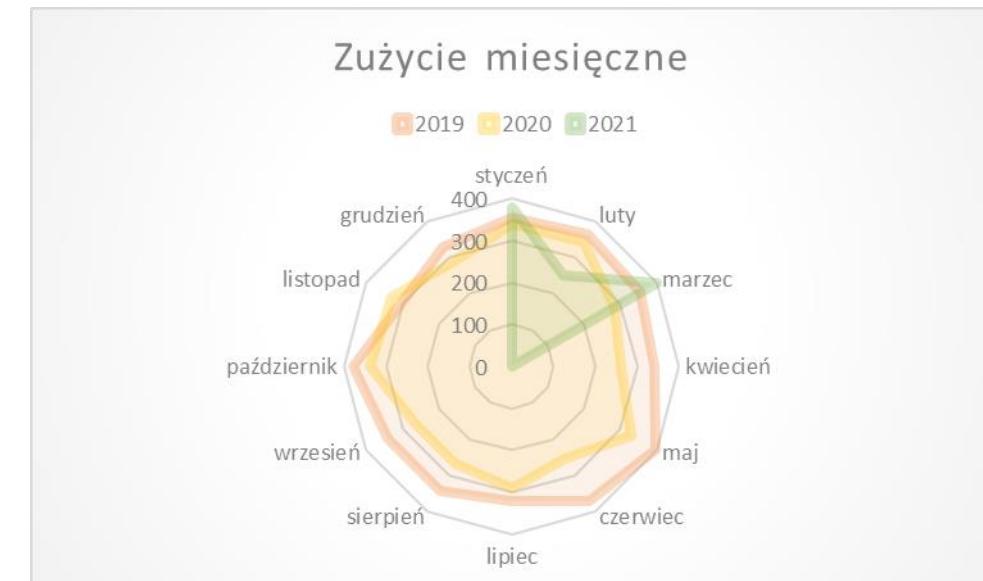
*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Wykrywanie trendu

- Fluktuacje danych mogą utrudniać „zobaczenie” trendu
- Możliwość dodania linii trendu
  - regresja liniowa
  - regresja wyższego rzędu
  - średnia ruchoma
  - wygładzanie
- Wzór funkcji regresji często nie daje żadnej informacji
- Średnia ruchoma wyliczana na podstawie określonej liczby poprzednich wartości

## Wykresy porównawcze

- Porównanie tej samej miary do wartości z poprzedniego analogicznego okresu
- Najważniejszy komunikat, który chcemy przekazać
- Uwagi:
  - im więcej okresów porównawczych, tym bardziej skomplikowana analiza
  - kluczowy dobór właściwego typu wykresu
  - można użyć wykresu 3D, jeśli czytelny



# Wykresy zawierające rozkład cech

- Rozkład cech widoczny na wykresie kołowym
- Niezbędna informacja o okresie, z którego pochodzą dane
- Wykres kołowy nie uwzględnia trendu
- Dobre praktyki:
  - dodaj etykiety danych
  - ogranicz liczbę kategorii do 10 (pozostałe 5-10% koła oznacz jako „inne”)
  - dla porównania przygotuj dwa wykresy (np. z różnych okresów)
  - sprawdź, czy wykres kołowy najlepiej prezentuje to, co ma być przekazane
- Czy potrzebna hurtownia danych, żeby przygotować dane do wykresu kołowego?

# Filtrowanie i tabele przestawne

- Filtrowanie
  - ograniczenie zakresu prezentowanego KPI
  - wygodne w przypadku raportów interaktywnych
  - przygotowanie określonych możliwości zmiany perspektywy
  - czy potrzeba porównania wykresów dla różnych filtrów?
  - możliwość analizy w głąb -> uwzględnianie kolejnych atrybutów
- Tabele i wykresy przestawne
  - porównanie danych dla różnych wartości tego samego atrybutu
  - możliwość analizy w głąb

# Podsumowanie

- Najważniejsze aspekty wizualizacji danych:
  - personalizacja dashboardu dla odbiorcy
  - wyświetlenie surowych danych w formie podsumowań
  - formatowanie warunkowe dla tabel zawierających podsumowania
  - dostosowanie formy do treści
- Wersja elektroniczna raportu nie musi zawierać zbyt wiele liczb
- Panel analityczny pozwoli na dotarcie do danych źródłowych uprawnionym użytkownikom
- Dodatkowe informacje na wykresie (adnotacje) ułatwiają zrozumienie zawartości



Fundusze  
Europejskie  
Wiedza Edukacja Rozwój



Politechnika Wrocławska

Unia Europejska  
Europejski Fundusz Społeczny



*„ZPR PWr – Zintegrowany Program Rozwoju Politechniki Wrocławskiej”*

# Hurtownie danych

Dziękuję za uwagę

dr inż. Marcin Maleszka