

СХОВИЩА ДАНИХ: Лекція №3

НУ “Львівська Політехніка”, кафедра ПЗ

Багатовимірна модель даних

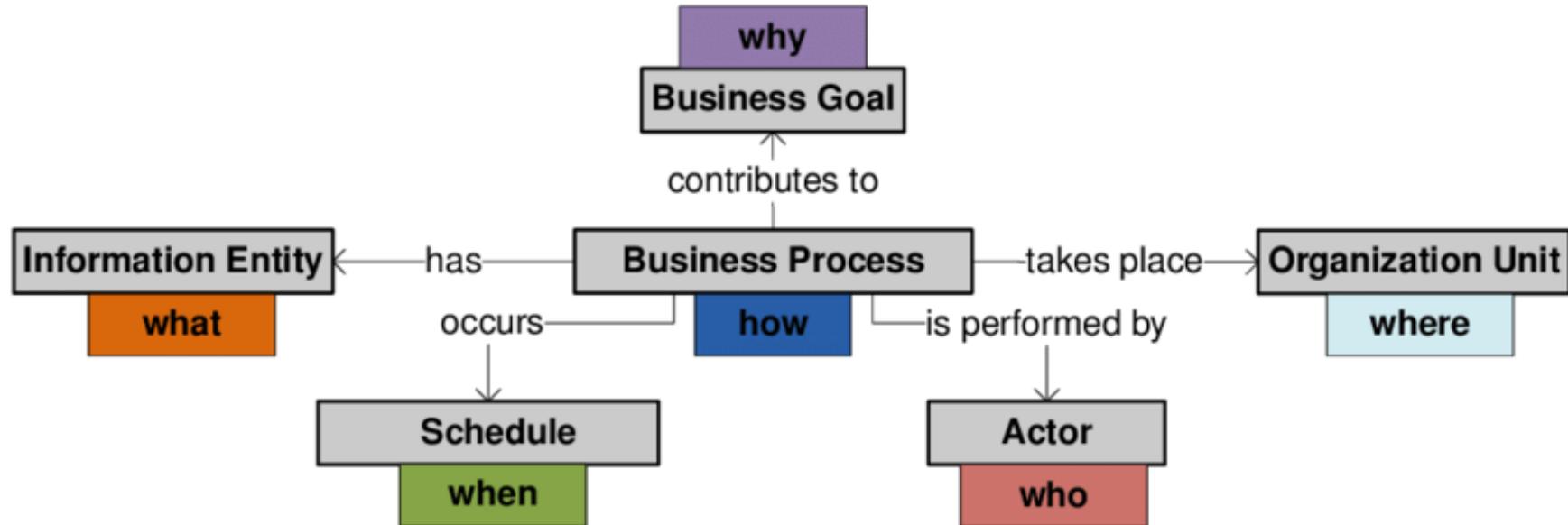
Моделювання сховища даних



- **Top-down** – повний збір всіх вимог для побудови єдиної схеми сховища даних
- **Bottom-up** – вітрини даних створюються окремо відповідно до вимог бізнес-аналізу
- **Analysis-driven** – аналіз вимог від потенційних користувачів різних рівнів організації
- **Source-driven** – структура наявних джерел даних задає особливості сховища

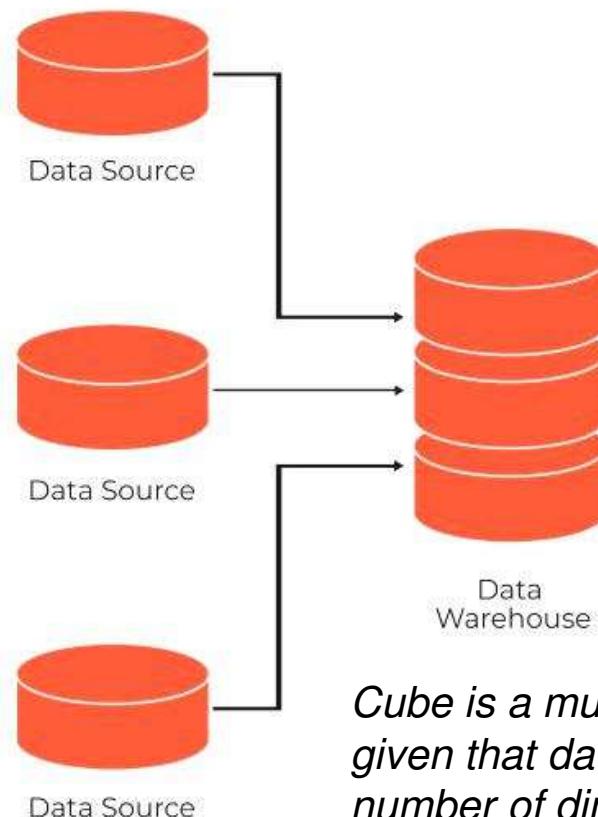
Від обліку бізнес-процесів до аналітичного представлення

- Опис процесів та структури даних, якими вони оперують
- Опис потоків даних, передумови потенційних змін у часі
- Інтерпретація зв'язків між сутностями предметної області
- Потреби аналітичної обробки на різних рівнях прийняття рішень
- Доцільність об'єднання даних різних бізнес-процесів
- Встановлення пов'язаних джерел для побудови цілісної моделі представлення характеристик різних сутностей
- Необхідність збереження та опрацювання історії змін

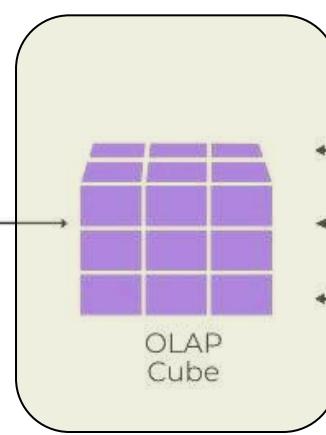


Аналітична обробка в реальному часі

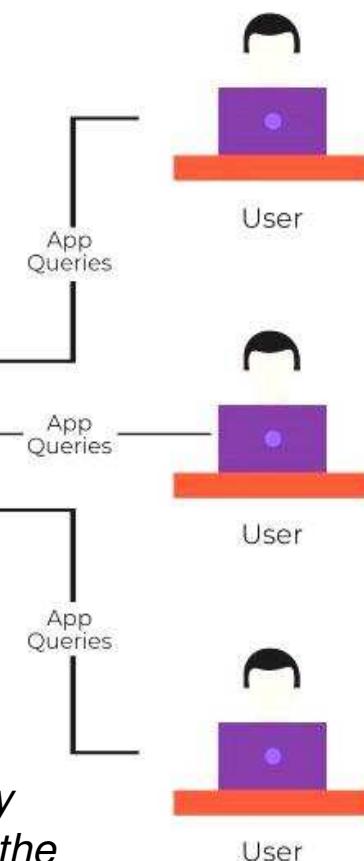
Консолідація



Агрегація



Представлення



Cube is a multidimensional dataset, given that data can have an arbitrary number of dimensions. Each cell of the cube holds a number that represents some measure of the business.

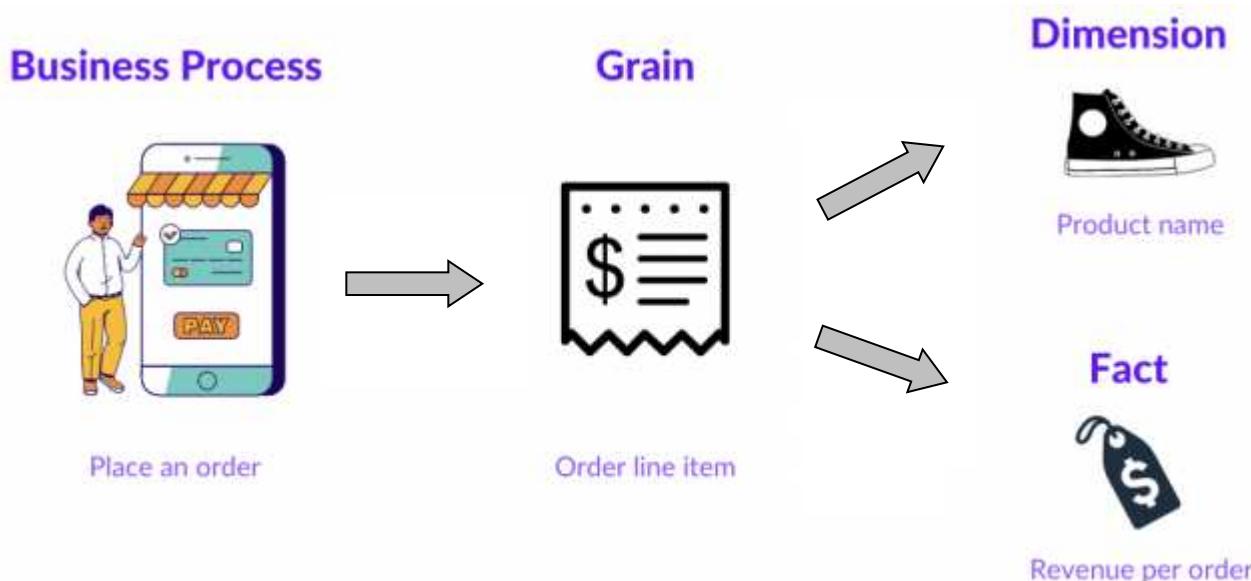
Обробка багатовимірних даних

- **Рівень представлення.** Інструменти візуалізації та маніпулювання, абстраговані від фізичної структури зберігання і методів накопичення
- **Рівень опрацювання.** Мова запитів та система обробки, яка формує вибірки даних за умовами
- **Рівень зберігання.** Фізична організація даних, орієнтована на ефективне виконання запитів. Використовується реляційна або інша модель

-
- ➔ **Технологія OLAP** – це сукупність вимог та засобів для швидкого різнопланового аналізу даних. Може бути реалізована навіть у простих електронних таблицях

Складові частини кубу

- **Гранулювання** визначає ступінь деталізації даних, достатню для вирішення аналітичних завдань
- **Виміри** задають контекст “*хто, що, де, коли, чому*” бізнес-процесу за допомогою описових атрибутів
- **Факти** зберігають корисні для аналізу “обставини”, фіксуючи цілісний результат подій в термінах вимірів
- **Метрика** є окремою числововою характеристикою факту, “вимірювальна” складова аналітичних звітів
- **Куб** зв’язує виміри і метрики факту в єдину сутність



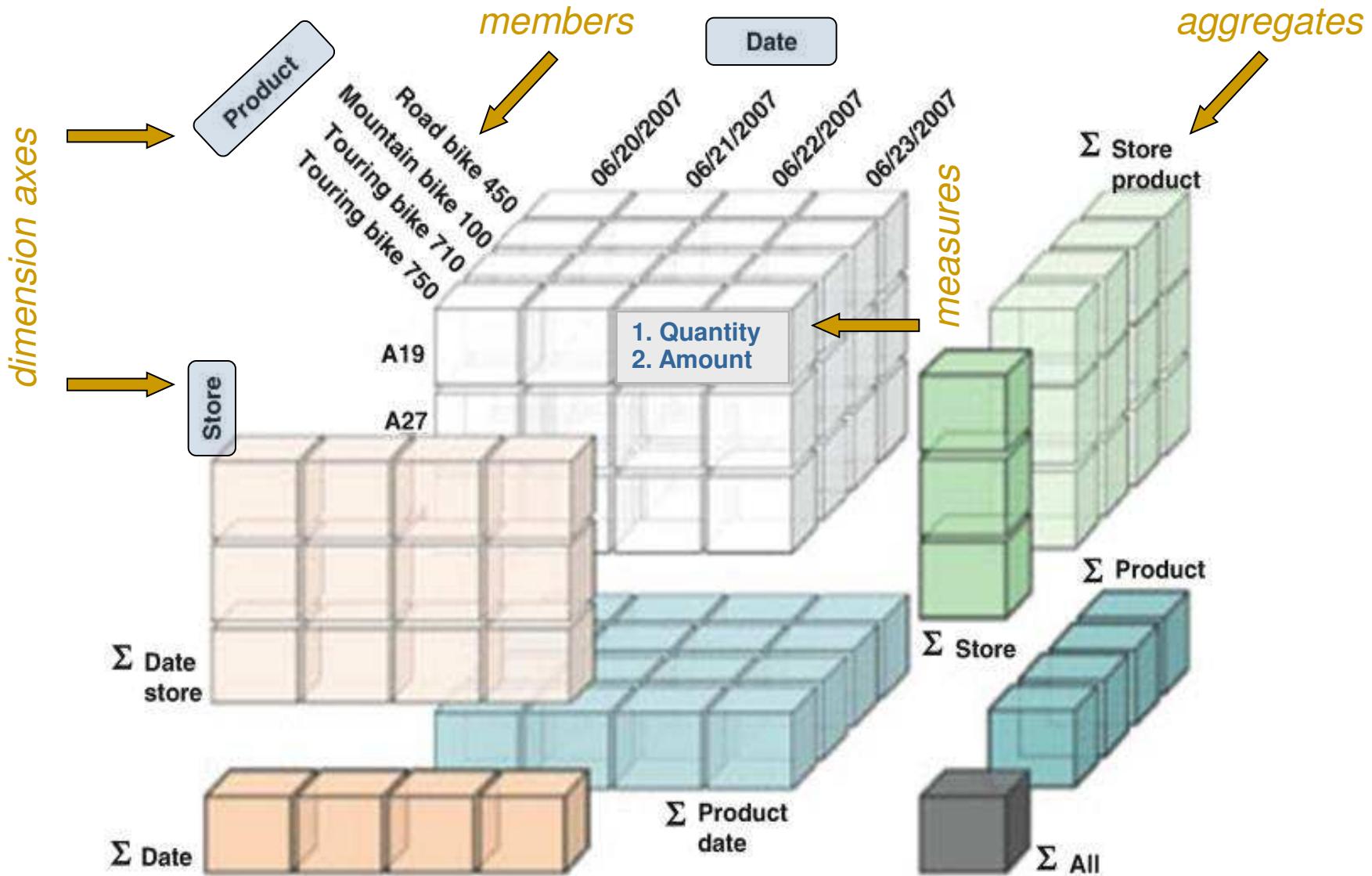
Гранулювання факту

- Рівень деталізації даних, що зберігаються у кубі, відповідно до потреб аналітики
 - транзакції або події – найнижчий рівень
 - періодичний – кумулятивна міра за певний час
- Факт-агрегат визначається сумою (або іншою функцією) значень метрик на більш високому рівні деталізації даних

On 01/01/2013 at 7:15, customer 0098745
bought product 12345 for the price of 10.95
EUR plus 20% VAT

On 01/01/2013, in our store “Brussels-av.
Louise”, 145 items of product 01245 have
been sold for an average price of 123.57 EUR

Об'ємна візуалізація кубу



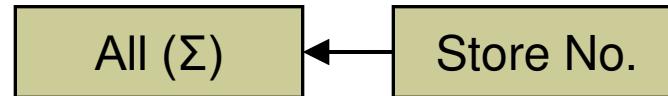
Таблична візуалізація кубу

		Measures							
		Quantity				Amount			
Dimensions		Store							
Product	Date	A19	A27	A34	Total Store	A19	A27	A34	Total Store
- Road Bike 450	06/20/2007	2	7	4	13	498	1743	996	3237
	06/21/2007	9	12	10	31	2241	2988	2490	7719
	06/22/2007	3		7	10	747		1743	2490
	06/23/2007	5	1	9	15	1245	249	2241	3735
	Total Road Bike 450	19	20	30	69	4731	4980	7470	17181
- Mountain Bike 100	05/20/2007	8	10	3	21	6392	7990	2397	16779
	05/21/2007	5	11	4	20	3995	8789	3196	15980
	05/22/2007	9	7		16	7191	5593		12784
	05/23/2007	6		4	10	4794		3196	7990
	Total Mountain Bike 100	28	28	11	67	22372	22372	8789	53533
- Touring Bike 710	05/20/2007	5		9	14	2995		5391	8386
	05/21/2007	7	2	12	21	4193	1198	7188	12579
	05/22/2007	4		13	17	2396		7787	10183
	05/23/2007		2	8	10		1198	4792	5990
	Total Touring Bike 710	15	4	42	62	9584	2396	25158	37138
+ Total Touring Bike 750		19	12	15	46	10621	6708	8385	25714
Total Product		82	64	98	244	47308	83765	49802	133566

Структура вимірів кубу

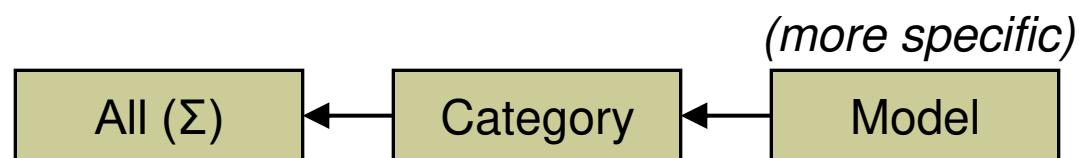
1. Store Dim.

- A19, A27, A34



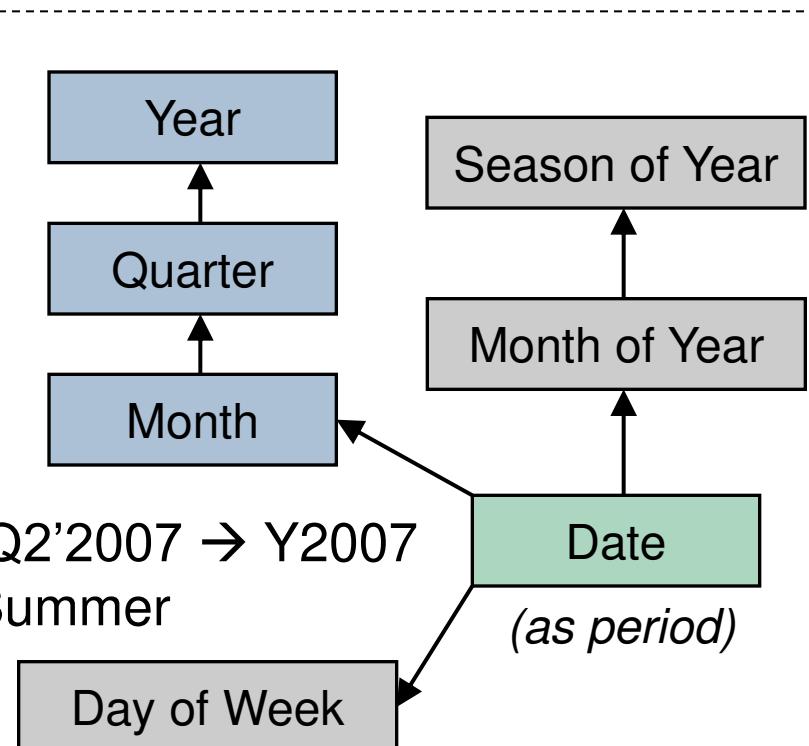
2. Product Dim.

- Road Bike
 - Model 400, 450, 500
- Mountain Bike
 - Model 100, 200
- Touring Bike
 - Model 710, 750, 800



3. Date Dim.

- 20.06.2007 → June 2007 → Q2'2007 → Y2007
- 20.06.2007 → 6th Month → Summer
- 20.06.2007 → Wednesday



Взаємодія користувачів із даними

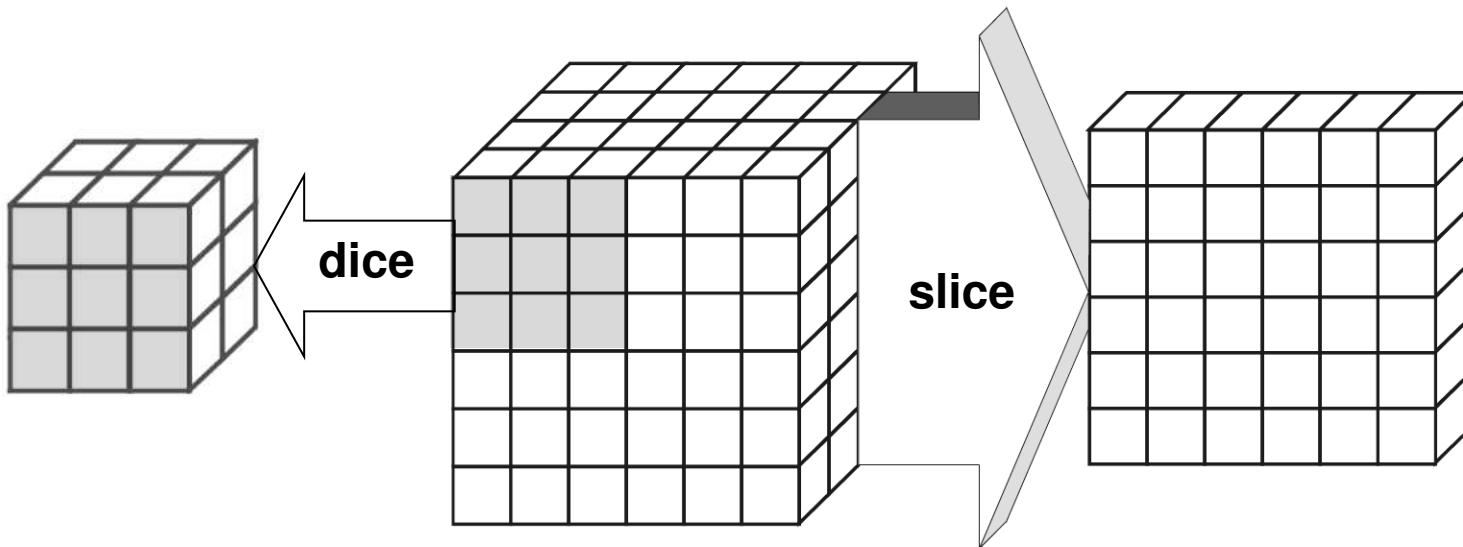
- Зручна проекція, агрегація, фільтрація, навігація по великих масивах даних
- Використання концепції з ієрархічними вимірами надає можливості для швидкого структурованого аналізу числової інформації
- Інтерактивний процес навігації відбувається за допомогою типових операцій на віртуальному кубі
- Графічна візуалізація даних в системах бізнес-аналітики допомагає виявляти тренди

*“Кожен клік – це крок до мети.
Кожен клік – це відволікання”*

(Ralph Kimball, 1996)

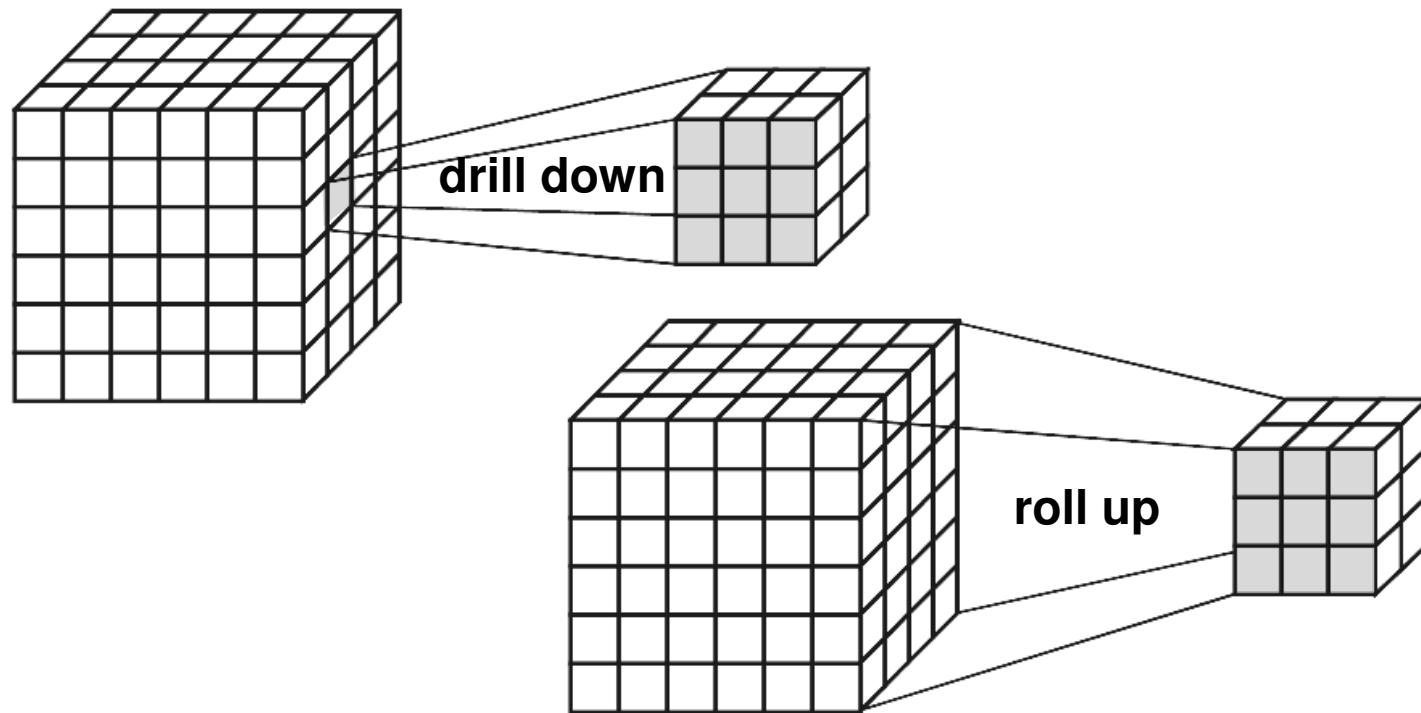
Отримання зрізу даних

- **Зріз** (slice – за одним виміром, dice – за декількома) обирає підмножину даних з кубу по заданим значенням елементів вимірів
- *Наприклад, “зріз даних за 2007 рік”*



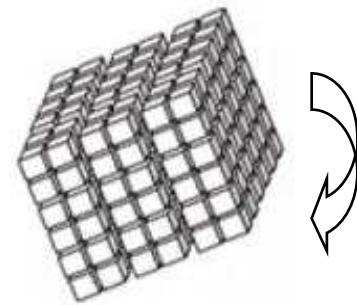
Керування деталізацією даних

- **Розгортання та згортання** (drill down, roll up) виконує навігацію за рівнями ієрархії вимірів з виведенням агрегатних значень метрик кубу
- *Наприклад, “перехід від років до рівня місяців”*



Інші операції з кубами

- Обертання** (pivot, rotate) - зміна порядку представлення вимірів, що застосовується при двомірному поданні даних
- Drill across – деталізація узгоджених даних з кількох кубів з однаковою структурою
- Drill through – деталізація від нижніх рівнів куба до даних джерел, з яких будується куб
- Scoping – обмеження представлення даних для конкретної групи користувачів
- Screening – обмеження екстракту даних
- Union – об'єднання декількох кубів, які мають однакову схему



Група	Кількість
■ Атрибутика	23
■ Брелоки	4
■ Значки	8
■ Килими для мишок	1
■ Магніти	3
■ Фігурки	7
■ Їжа	1
■ Напої	1
■ Одяг	46
■ Головні убори	2
■ Футболки	26
■ Худі, світшоти	3
■ Шкарпетки	15
■ Поліграфія	3
■ Листівки	3
■ Посуд	4
■ Кружки	4

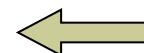
Приклади операцій з кубом продажів

Вимір: "Група"

Метрика: "Кількість"

Група	Кількість
■ Брелоки	4
■ Значки	8
■ Килими для мишок	1
■ Магніти	3
■ Фігурки	7
Σ	23

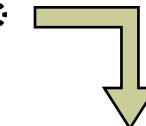
розгорта



Група	Кількість
■ Атрибутика	23
■ Їжа	1
■ Одяг	46
■ Поліграфія	3
■ Посуд	4
Σ	77



зріз

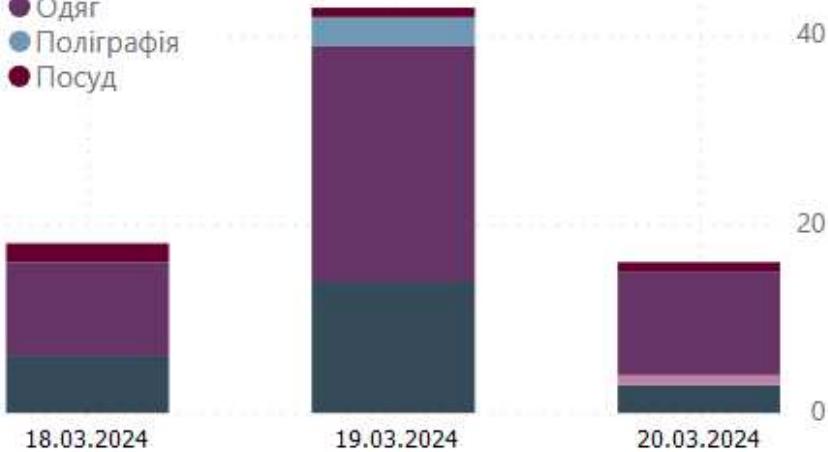


деталізація

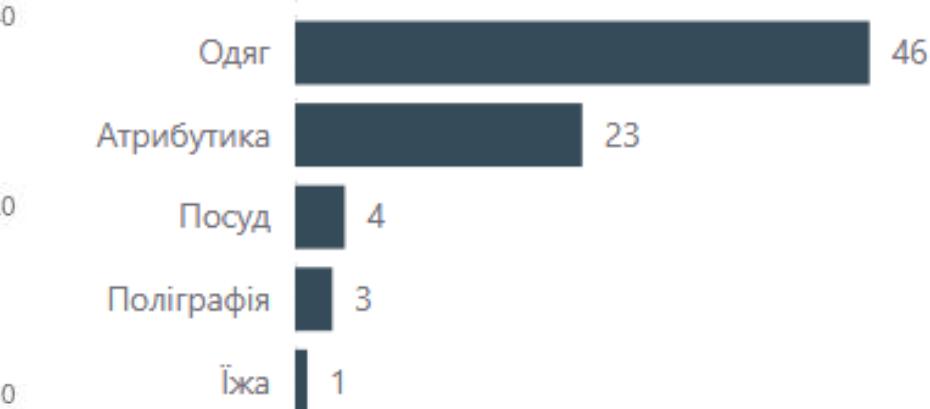
Група	Назва	Рік	Місяць	День	Сума	Кількість
Посуд	Кружка BU - Harry Potter - Slytherin	2024	Март	18	290,00	1
Посуд	Кружка BU - Shrek	2024	Март	19	290,00	1
Посуд	Кружка BU - The Office	2024	Март	18	290,00	1
Посуд	Кружка RWB - Linkin Park - Minutes Midnight	2024	Март	20	240,00	1
Σ					1110,00	4

Приклади візуалізації продажів по вимірах

- Атрибутика
- Іжа
- Одяг
- Поліграфія
- Посуд



Дата + група

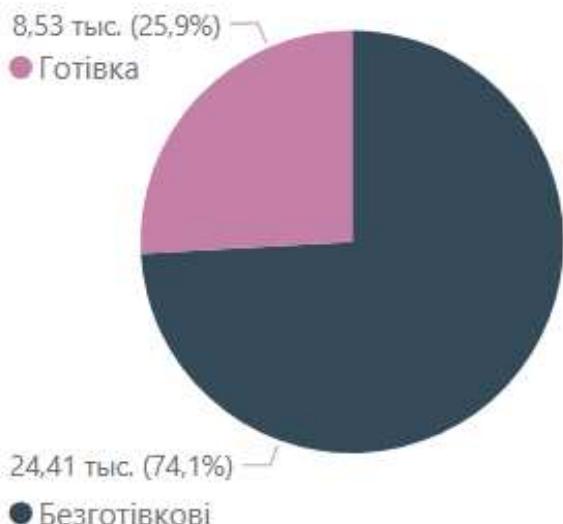


Група (верхній рівень)

- Анімація
- Музика
- Культура
- Кіно
- Ігри



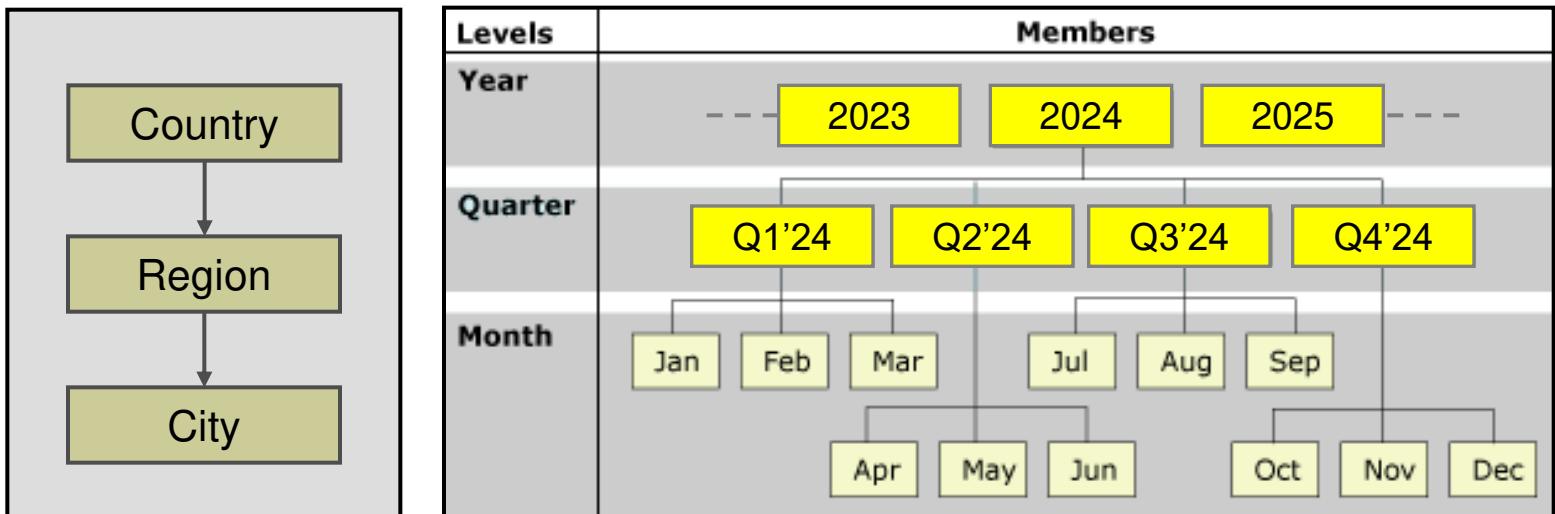
Спосіб оплати



Напрямок + тема

Ієрархія вимірів

- Організовує елементи у деревовидну структуру, що дозволяє змінювати ступінь деталізації даних для аналізу
- Семантичний зв'язок між елементами різних рівнів “частина – ціле”, “батько – нащадок”, “детальне – загальне”



- Рівень задає глибину вкладеності елементу в гілці ієрархії
- Ієрархія розбиває елементи виміру на окремі підмножини
- Паралельні ієрархії забезпечують можливість різного “погляду” на один факт в контексті одного виміру

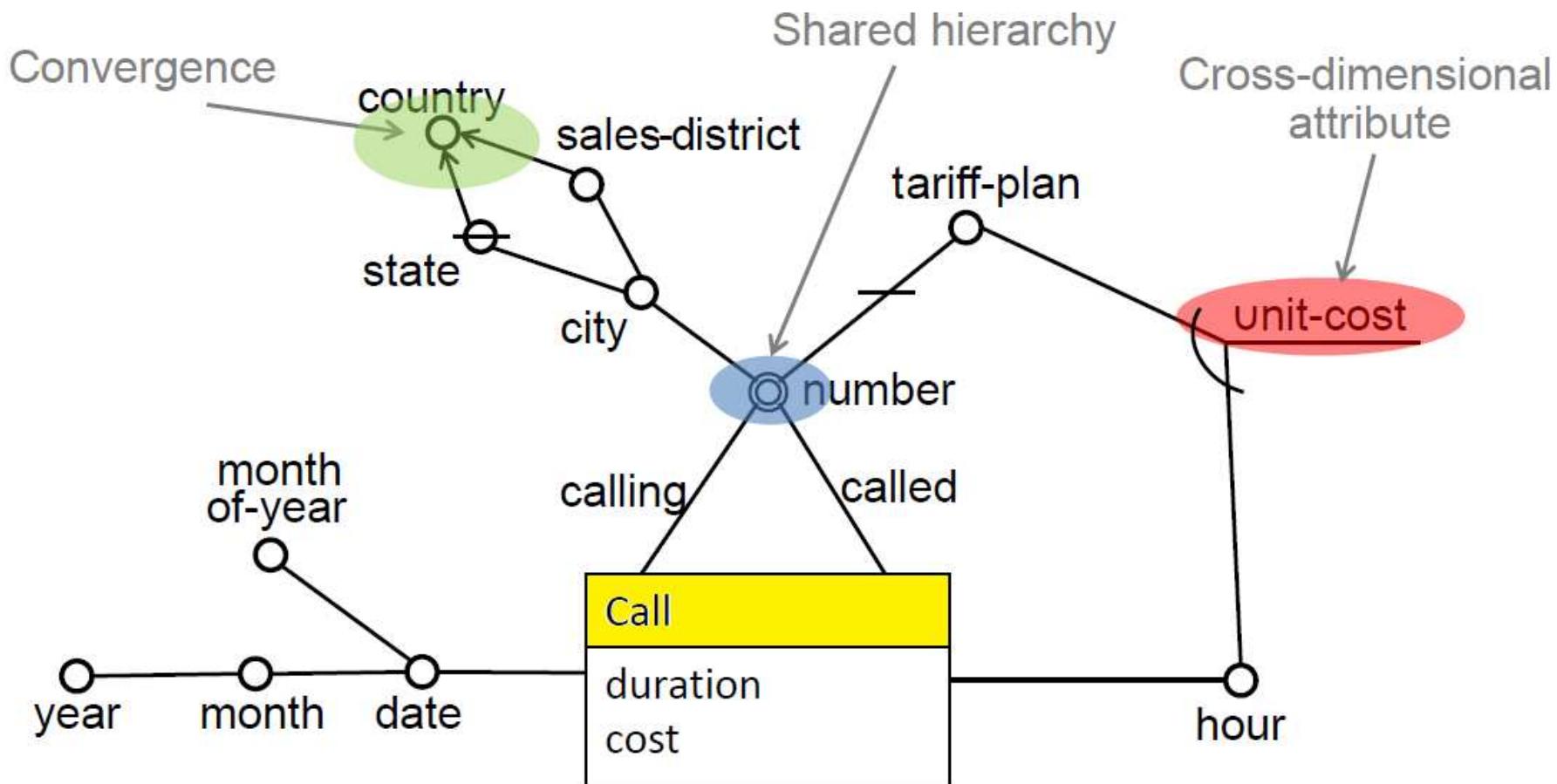
Властивості ієрархії вимірів

- **Збалансована** – всі гілки мають однакову кількість рівнів
 - дата: *рік – квартал – місяць – день*
 - товар: *країна – бренд – назва*
- **Незбалансована** – різні гілки мають різну кількість рівнів
 - підрозділи: *відділ – група – керівництво*
 - вкладені категорії товару: *батько – нащадок*
- **Нерівна** (*incomplete*) – деякі рівні можуть бути пропущені при визначеному більш загальному
 - адміністративно-територіальний устрій:
область – район – місто – квартал
- **Множинна** (*multiple*) – декілька елементів нижнього рівня відповідають одному верхньому
 - автори: *книжка – автор*

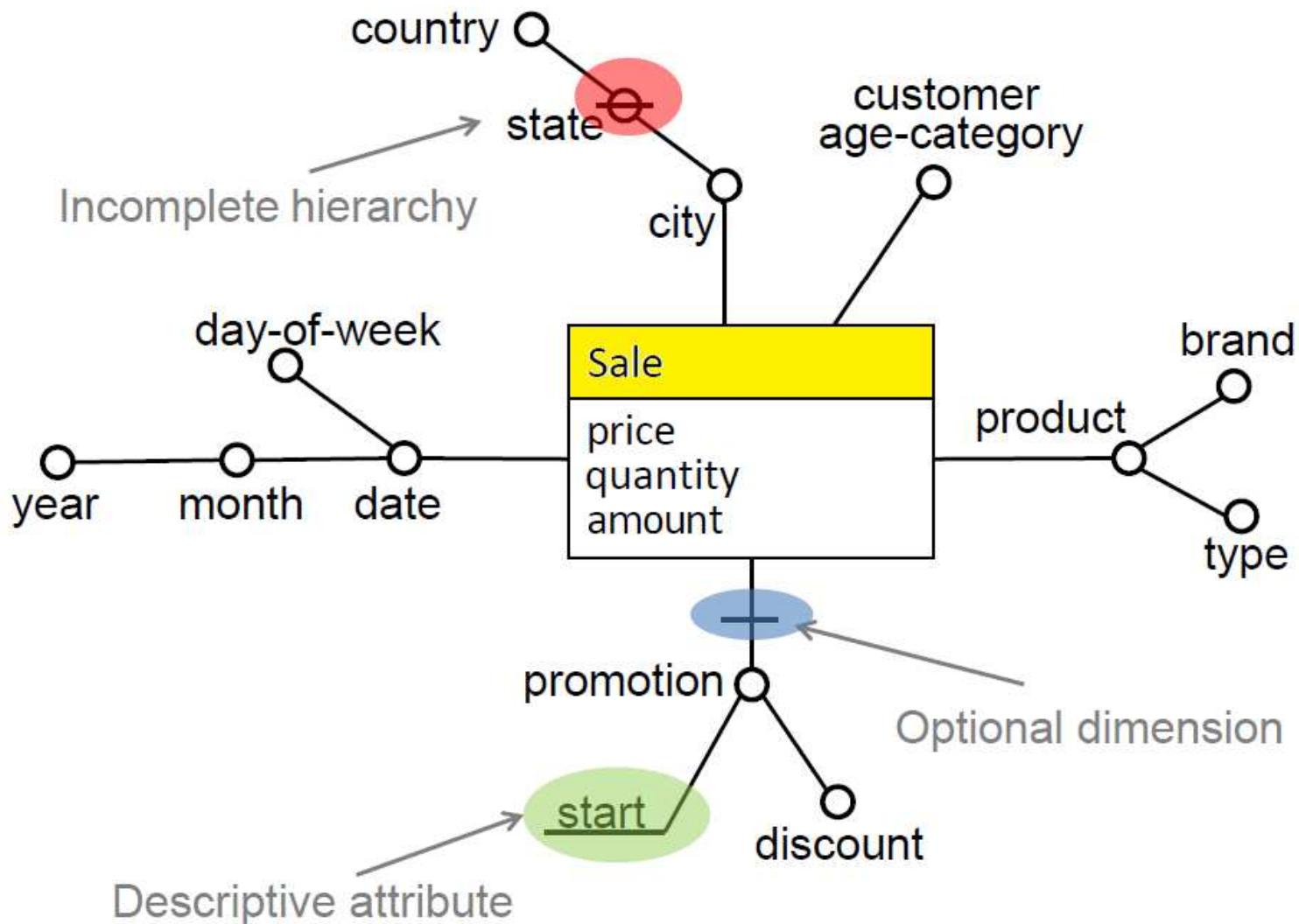
Особливості атрибутів вимірів

- **Повільно змінні** (slowly changing) – атрибути можуть міняти значення із часом, з'являється проблема збереження історії
 - при оновленні встановити нове значення атрибута
 - створити новий рядок в таблиці вимірів
 - додати атрибут в таблицю з новим значенням
- **Швидкозмінні** (rapidly changing) – опис деяких атрибутів змінюється за короткі проміжки часу
 - задати діапазони дискретних значень виміру
 - розбити на декілька вимірів
- **Вироджені** (degenerate) – не має жодних атрибутів, тому не потребує окремої таблиці, часто є ідентифікаторами зовнішніх систем
 - надає групування споріднених записів фактів
 - підтримує зв'язок із джерелом даних

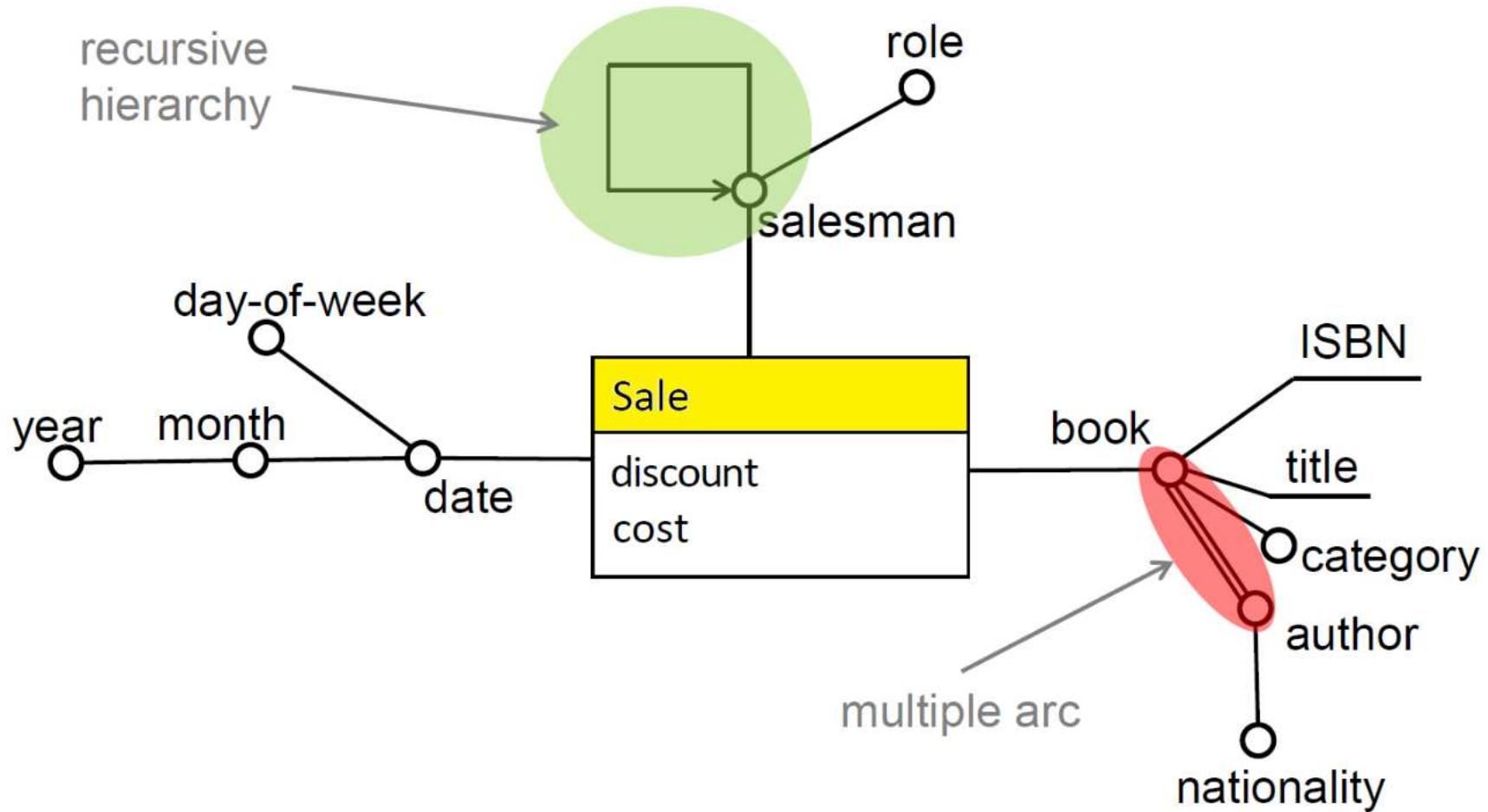
Приклад факту: спільні значення



Приклад факту: необов'язковість даних



Приклад факту: складна ієрархія



Типи метрик фактів

- **Адитивні** (additive) – можливе використання з будь-якими вимірами для підсумовування
 - кількість замовлень, товарів у чеку, сума прибутку
- **Напівадитивні** (semiadditive) – можливе підсумовування з деякими вимірами
 - залишок на рахунку, час обслуговування клієнта
- **Неадитивні** (non-additive) – підсумовування не має сенсу або неможливе
 - ціна товару, відсоток знижки, вік користувача
- **Міри інтенсивності** (measures of intensity) – насиченість показника в залежності від масштабу виміру, зазвичай розраховується як відношення агрегованих значень інших метрик
 - середній чек = об'єм продажів / кількість замовлень
 - рентабельність = чистий прибуток / об'єм продажів

Типи таблиць фактів

- Таблиця **транзакцій** містить характеристики завершеної події на встановлений момент часу
 - *купівля товару, виплата постачальнику, нарахування боргу*
- Таблиця **періодичних** моментальних знімків збирає дані, що фіксують стан певного напряму на певний час
 - *звернень в годину, витрати за день, відкрито замовлень*
- Таблиця **кумулятивних** моментальних знімків накопичує підсумки прогресивних показників події на певний час
 - *прибуток з початку року, баланс рахунку, поточні запаси*
- Таблиця **агрегатів** створюється для швидкої обробки дрібної грануляції факту на більш загальному рівні або коли певний вимір взагалі не потрібен
 - *продажі по підрозділах за рік, щомісячні суми податків*
- Таблиця **відстеження** подій фіксує факти, які не мають числових метрик, але встановлені час та опис
- Таблиця **охоплення** подій зберігає інформацію про активності за певний період, визначений вимірами. Враховується можливе перекриття у часі

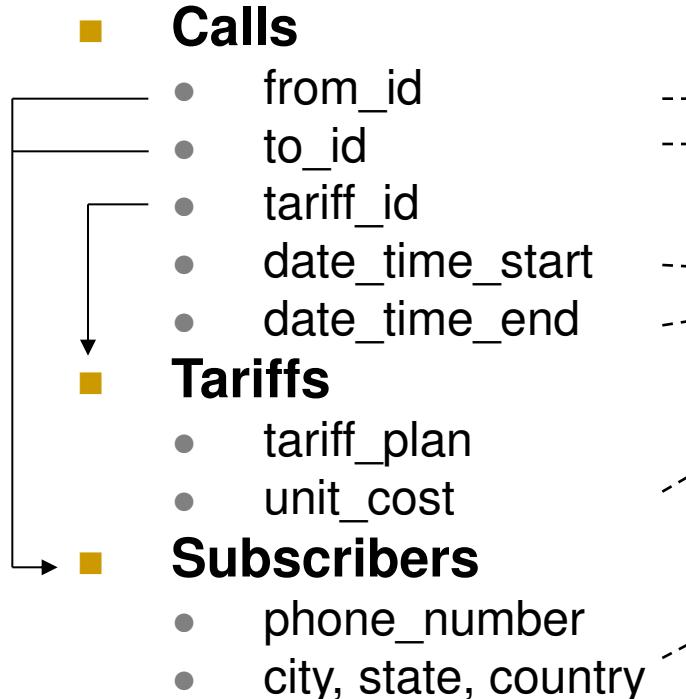
Приклад транзакційного факту



OLTP-дані – таблиці реєстрації дзвінків між двома абонентами



OLAP-куб – аналіз найбільш прибуткових напрямків дзвінків



Transactional Data

- from_id
- to_id
- tariff_id
- date_time_start
- date_time_end

Measures

- calling_number
- called_number

Dimensions

- tariff_plan
- from_city → state → country
- to_city → state → country
- date → month → year
- hour = hour (time_start)



Приклади аналізу даних

■ По транзакціях

- calling_number
- called_number
- tariff_plan
- hour
- date
- cost
- duration

} *єси вимірю, низькій рівень*

■ По датах в групах

- date
- month
- year
- cost
- duration
- count

} *1 вимір, єси рівні*

■ По годинах доби

- tariff_plan
- hour
- cost
- duration

■ По напрямках

- from_city
- to_city
- cost

From To	City1	City2	City3
City1			
City2			
City3			

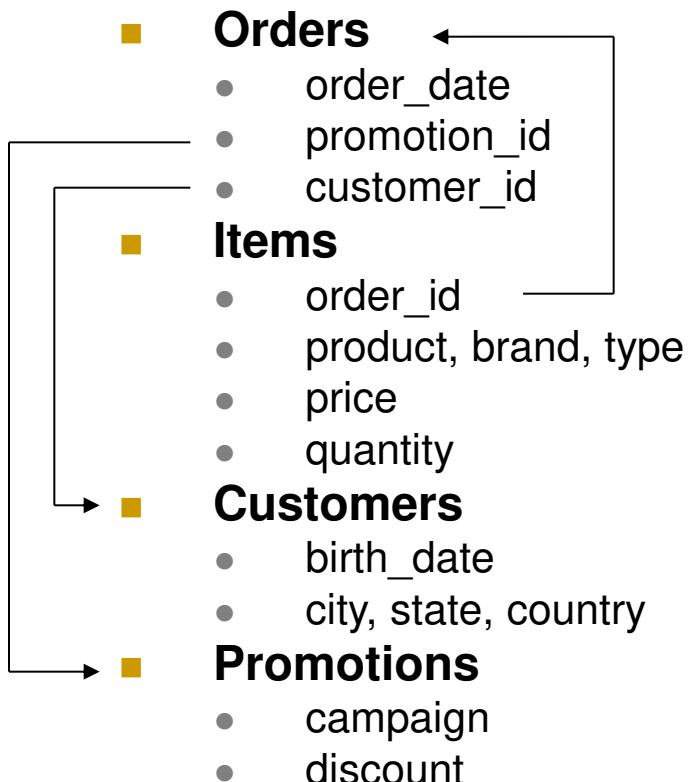
Приклад кумулятивного факту



OLTP-дані – таблиці обліку виконаних замовлень товарів



OLAP-куб – аналіз продажів за категоріями клієнтів



Aggregated Data

- order_date
- product
- customer_city
- customer_age

Measures

- price = avg (item_price)
- quantity = sum (item_quantity)
- amount = sum (price * quantity)



Dimensions

- date → month → year
- city → state → country
- age_cat = age (birth_date) / 10
- product → type
- product → brand
- promotion → discount

Приклади аналізу даних

■ По віку клієнтів

- age_category
- year
- count

■ По країнах клієнтів

- customer_country
- month
- discount
- amount

3-й етап роботи з даними
таблицю

Country		UA	UK	US
Month				
Jan 2024	0%			
	5%			
	10%			
Feb 2024	0%			
	5%			
	10%			

■ По кожному продукту

- date
- product
- price

■ По типу продуктів

- product_type
- customer_country
- amount
- quantity

■ По бренду продуктів

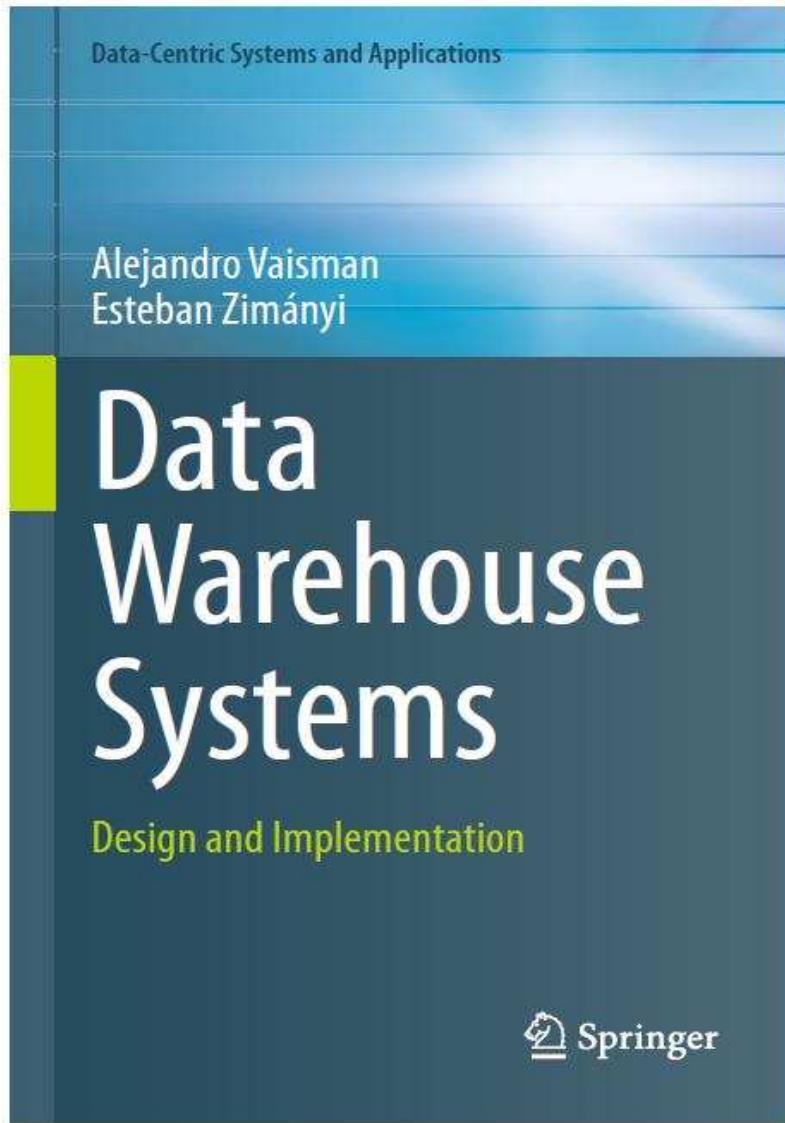
- year = (2024)
- product_brand
- age_category
- amount
- quantity

середня ціна
на дату

зріз по
етапу даних

- *Fact*: most specific unit of data that will be used in the analysis.
 - Usually corresponds to one or more transactions within a company.
- *Dimension*: A fact property; a coordinate of the fact.
 - Every fact corresponds to a unique combination of values for the dimensions.
- *Measure*: Numerical property of a fact; describes a quantitative aspect relevant for the analysis.
 - Measures can be aggregated, grouping by the dimensions, using an aggregation function to form secondary events.
- Dimensions and hierarchies define how data can be aggregated.

Дякую за увагу!



The Data Warehouse Toolkit

Third Edition

The Definitive Guide
to Dimensional
Modeling

Ralph Kimball
Margy Ross



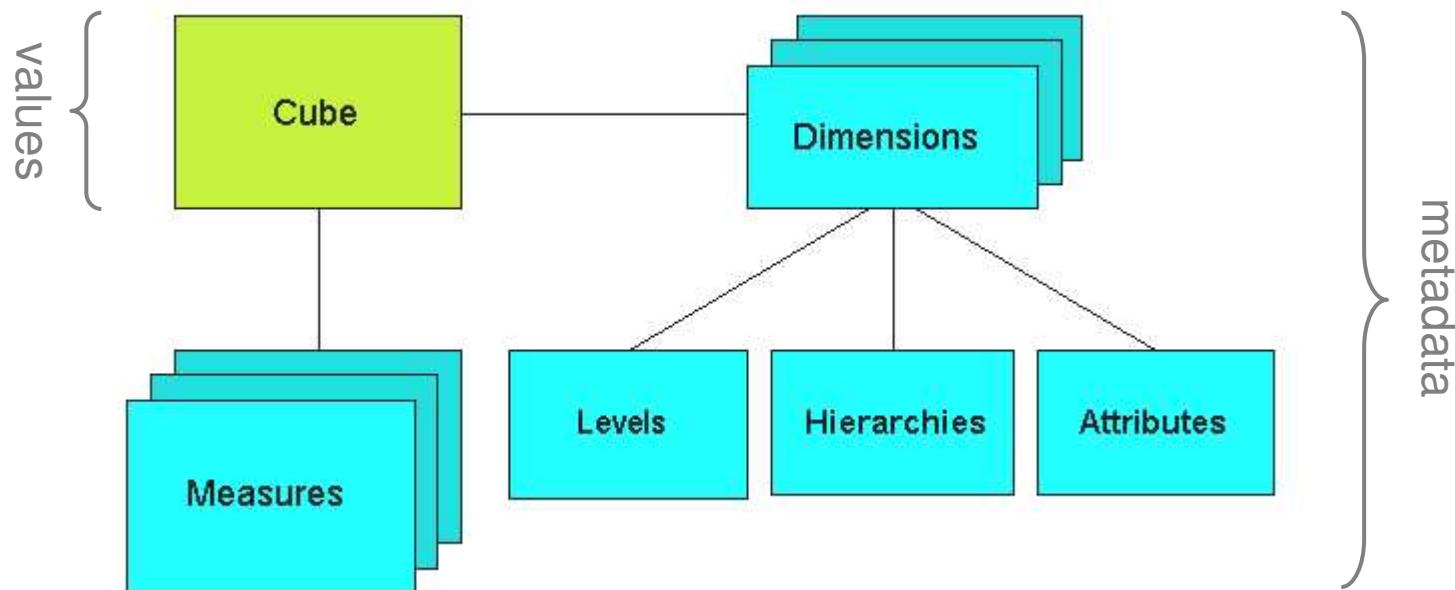
СХОВИЩА ДАНИХ: Лекція №4

НУ “Львівська Політехніка”, кафедра ПЗ

Реляційна модель сховища даних

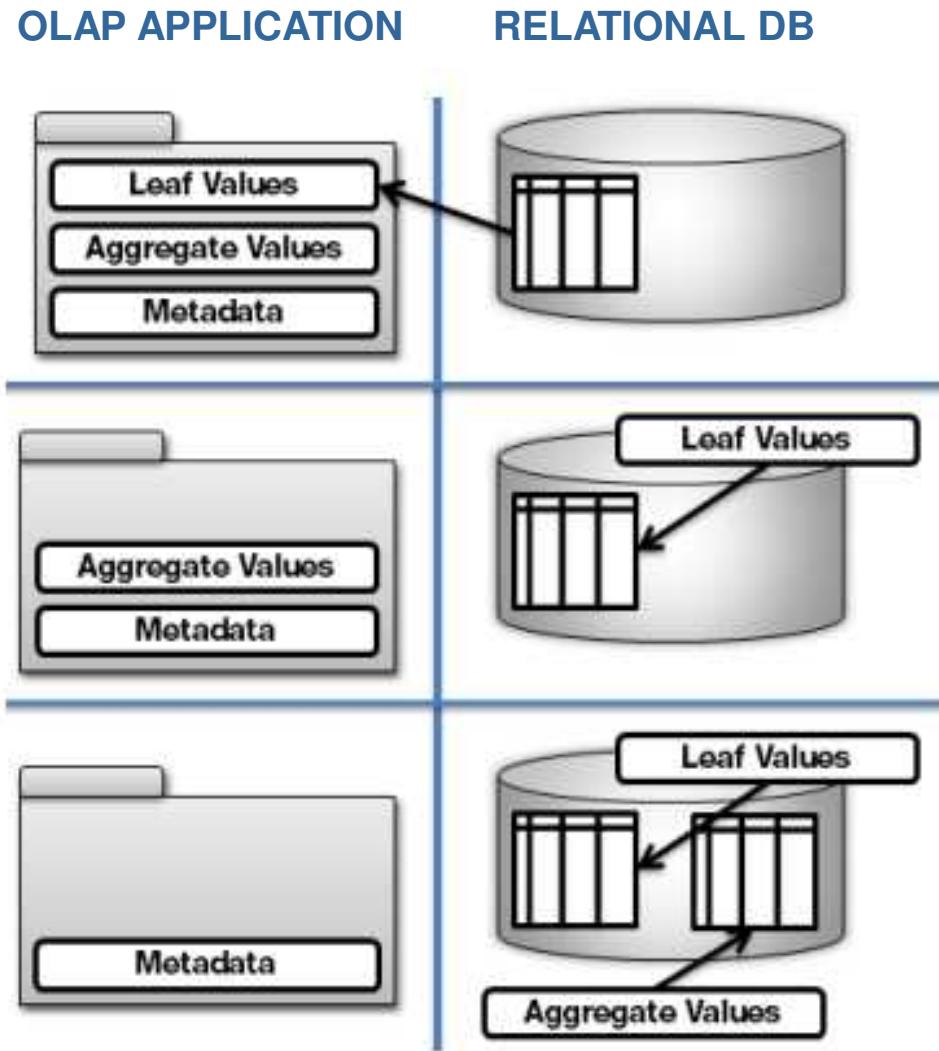
Інформація у сховищі даних

- **Куби** зберігають числові показники та характеристики завершених бізнес-операцій
- **Виміри** містять набори елементів, описаних символічними даними, які були актуальні на момент часу відповідної операції
- **Метадані** визначають ієрархії вимірів, властивості метрик, стан та особливості внесення даних, а також додаткові аспекти функціонування сховища

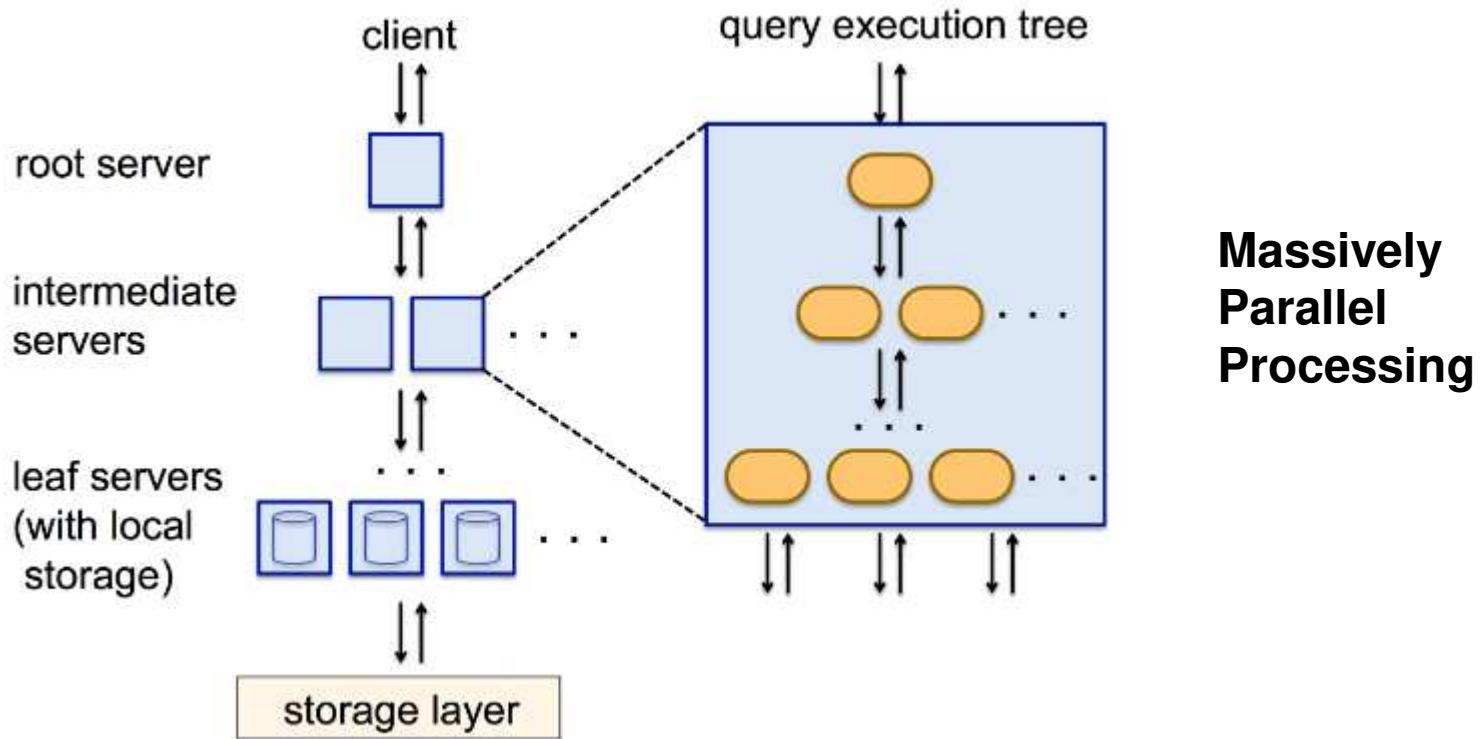


Класичні підходи реалізації сховища даних

- **MOLAP (Multidimensional)**
 - дані зберігаються в окремій БД у формі багатомірних масивів
 - наперед обчислені значення за рівнями вимірів
 - спеціальні операції внесення та вибірки даних
- **HOLAP (Hybrid)**
 - багатовимірна БД тільки для агрегованих значень
 - таблиці для даних максимальної деталізації
- **ROLAP (Relational)**
 - всі дані зберігаються в таблицях реляційної БД
 - метадані описують багатовимірне подання
 - структурована мова запитів, гнучкість вибірки даних
 - додаткові технології швидкого з'єднання та агрегації



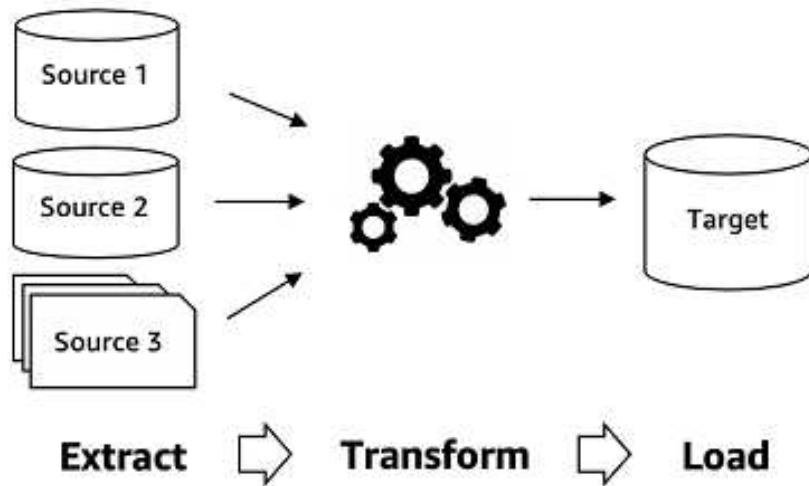
Сучасний підхід на основі паралельних обчислень



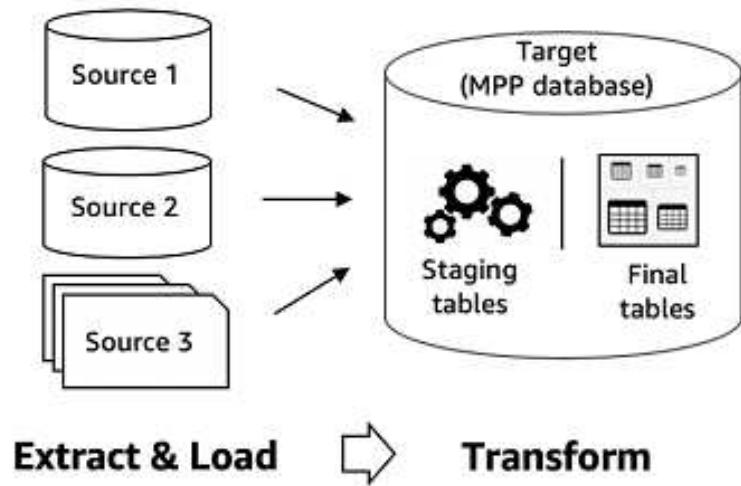
- Швидка реалізація необхідних вітрин даних
- Замість складного ETL-процесу (трансформація перед завантаженням) проектується ELT-процес (накопичення даних в озерах)

Порівняння підходів організації даних

- Трансформація даних операційних джерел у цільовий формат сховища
- Запити до цілісної аналітичної бази



- Дані у сховище завантажуються у первинному вигляді
- Трансформація за потребою аналізу
- Окремі куби даних



Програмне забезпечення MOLAP

- Зберігає метадані з описом OLAP-кубу
- Підтримує багатовимірні запити мовою MDX
- Спеціальний інтерфейс для внесення даних
- Власні технології зберігання даних кубу
 - Microsoft SQL Server Analysis Services
 - Oracle Essbase
 - IBM Cognos TM1
 - SAP Business Warehouse
 - Palo OLAP Server (open source multidimensional storage)
- Запити транслюються до табличного сховища
 - MicroStrategy Intelligent Cubes
 - PARIS Technologies Olation
 - Board Enterprise Planning Platform
 - Pentaho Analysis Services (based on Mondrian engine)
 - Apache Kylin (open source warehouse platform)

Програмне забезпечення ROLAP

- Реляційна СУБД для зберігання даних
- Мова запитів SQL оперує таблицями
- Реалізовані спеціальні типи індексів
 - Microsoft SQL Server
 - Oracle Database (індекси bitmap)
- Оптимізація фізичного рівня зберігання даних
 - SAP HANA Database (предобробка агрегатів)
 - Vertica, Actian Vector, MonetDB (векторні БД)
 - Greenplum (на базі PostgreSQL)
 - ClickHouse, Exasol, Apache Druid
- Хмарні рішення використовують паралельні обчислення для пришвидшення агрегації
 - Amazon Redshift
 - Google BigQuery
 - Azure Synapse Analytics
 - Snowflake

Корпоративна модель даних

■ Включає різноманітний опис:

- предметної області (областей) організації
- структур даних предметних областей
- бізнес-процесів і бізнес-процедур
- потоків даних, прийнятих в організації

■ Розглядає предметні області:

- різні аспекти діяльності організації і з різним ступенем деталізації і завершеності
- групи сущностей, які відносяться до підтримки конкретних потреб бізнесу

Модель сховища даних

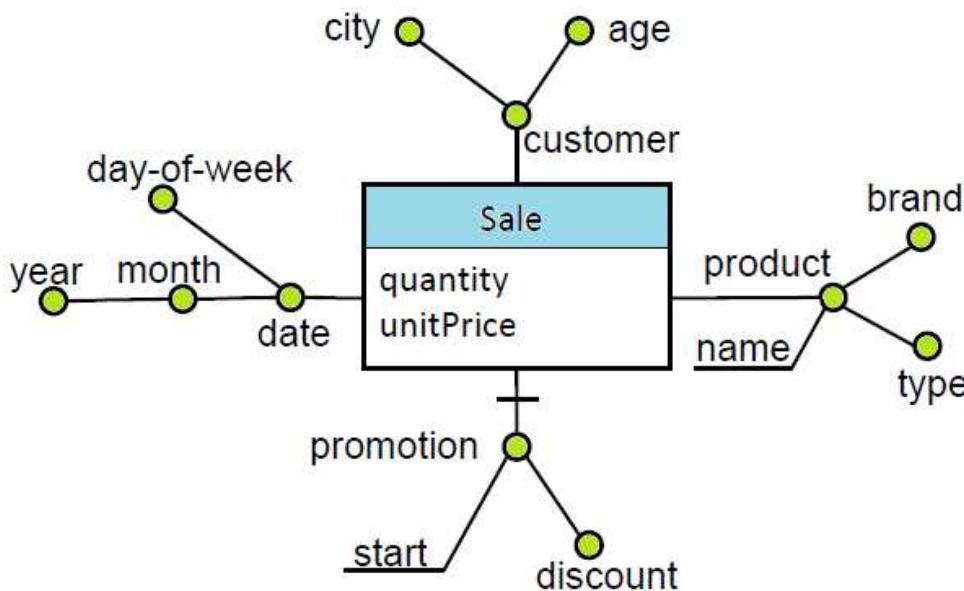
- Алгоритм побудови сховища даних на основі корпоративної моделі:
 - дослідити часові залежності даних і, якщо необхідно, додати елемент часу в ключі сутностей СД
 - додати у модель похідні (обчислювані) елементи даних
 - перетворити зв'язки між сутностями
 - визначити рівень структуризації (гранулювання) даних у СД
 - сформувати логічну схему СД з таблиць корпоративної моделі даних

Відбір даних для сховища

- Врахування можливих змін даних з часом
- Збереження історії певних характеристик
- Похідні метрики процесів є статичними в розрізі визначених вимірів
- Не слід включати:
 - дані, час життя яких в КМД дуже малий з точки зору часових масштабів СД
 - дані, що не входять у тимчасові залежності, які зберігаються в СД
 - дані, які мають значення або сенс тільки при оперативній обробці даних

Проектування реляційної моделі

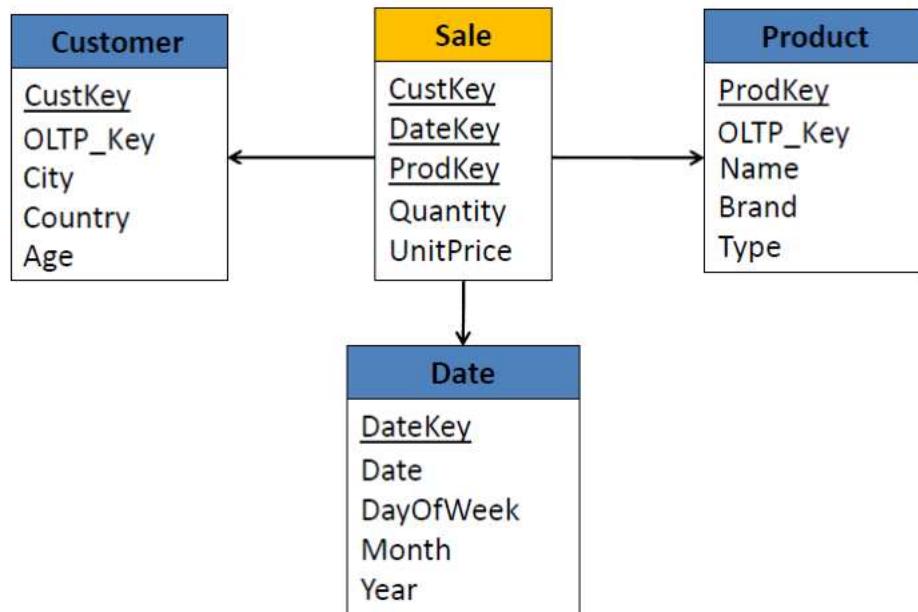
- Встановлення об'єктів та зв'язків, що описують багатовимірний куб в термінах бізнес-користувачів
- Створення схеми реляційної БД, будова якої відповідає певним правилам з урахуванням структури інформації предметної області



Feature	Conceptual	Logical	Physical
Entity Names	✓	✓	
Entity Relationships	✓	✓	
Attributes		✓	
Primary Keys		✓	✓
Foreign Keys		✓	✓
Table Names			✓
Column Names			✓
Column Data Types			✓

Вимоги до реляційної моделі

- Швидка вибірка, повільне оновлення
- Таблиці фактів та вимірів, таблиці-мости
- Компактні реляційні ключі, індекси на атрибути
- Попередні значення атрибутів (історичні дані)
- Значення бізнес-ключів з OLTP-бази



- Кешування агрегатів на обраних рівнях
- Bitmap-індекси для ключів вимірів
- Колонкове зберігання таблиці фактів
- Розділення таблиць на партиції

Типи схеми бази даних для багатовимірної моделі

- “Зірка” (star) має одну таблицю фактів і по одній денормалізованій таблиці на кожний вимір
- “Сніжинка” (snowflake) має одну таблицю фактів і декілька нормалізованих таблиць вимірів
- “Сузір’я фактів” (fact constellation, galaxy) має декілька таблиць фактів, пов’язаних реляційними ключами до спільних таблиць вимірів

Властивості таблиць фактів

- Мають сурогатний первинний ключ для транзакційного факту
- Або мають складений ключ із зовнішніх ключів до таблиць вимірів
- Відсутня функціональна залежність між ключами вимірів
- Поля із числовими метриками використовуються для агрегації
- Додаткові колонки мають службовий або інформативний характер
- Вибір вимірів та метрик обумовлений потребами бізнес-аналітики та міркуваннями ефективності
- Один багатовимірний куб може мати декілька таблиць фактів-агрегатів

factSales	
PK	SaleID
FK1	DateKey
FK2	ProductKey
FK3	StoreKey
FK4	CustomerKey
	SalesQuantity
	SalesPrice
	SalesAmount
TK	ReceiptID

aggSales	
FK1	DateKey
FK2	ProductKey
FK3	StoreKey
	SalesQuantity
	SalesAmount

Властивості таблиць вимірів

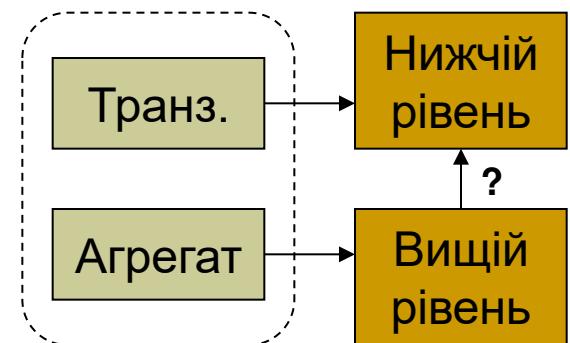
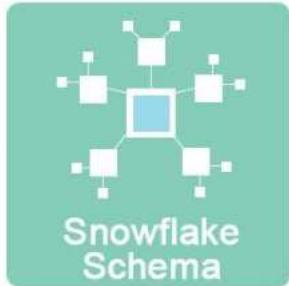
- Містять дані про деталізацію фактів
- Можуть мати додаткову інформацію для опису ієархії
- Компактний первинний ключ, незалежний від операційних даних
- Ступінь нормалізація таблиць залежить від обраної схеми та міркувань ефективності
- За значенням колонок формуються звіти та запити
- Дані можуть змінюватись із часом, при цьому зберігається зв'язок із старими записами таблиць фактів
- Можуть мати колонки OLTP-ключів для реалізації синхронізації

dimDate	
PK	DateKey
	FullDate
	Year
	Month
	Day
	DayOfWeek

dimStore	
PK	StoreKey
	Name
	Street
	City
	Country
TK	StoreID

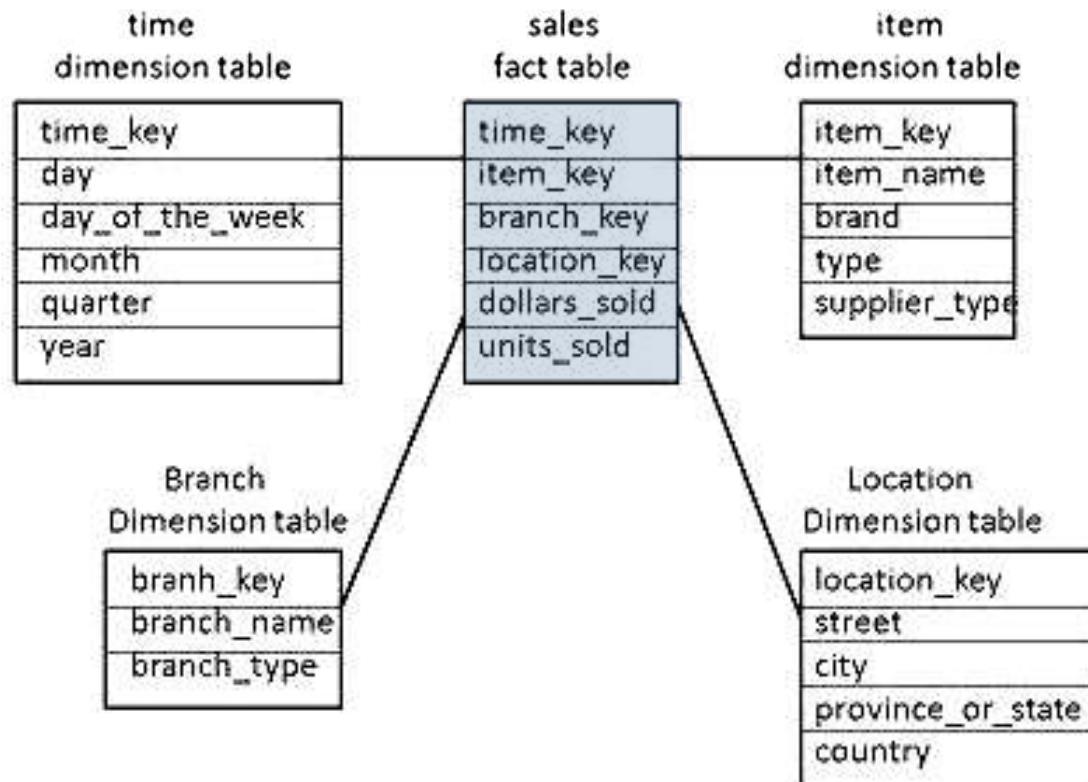
Нормалізація таблиць вимірів

- Широкі денормалізовані таблиці швидші на вибірку, мають комплексні індекси, зручні сурогатні ключі
- Але потребують більше місця при багатократному дублюванні значень, повільне внесення даних
- Дані у сховищі оновлюються автоматичними системами, тому можливі аномалії та час не є суттєвими
- Чим більша кількість таблиць на 1 вимір, тим складніші процеси оновлення з OLTP-джерел даних
- Структура ієархії, особливості характеристик, обсяг даних визначають потреби в нормалізації та декомпозиції
- Додаткові факти-агрегати, визначені на інших рівнях грануляції, потребують зовнішніх ключів до окремої таблиці
- Класичний ER-підхід поступається місцем вимогам аналітичних запитів



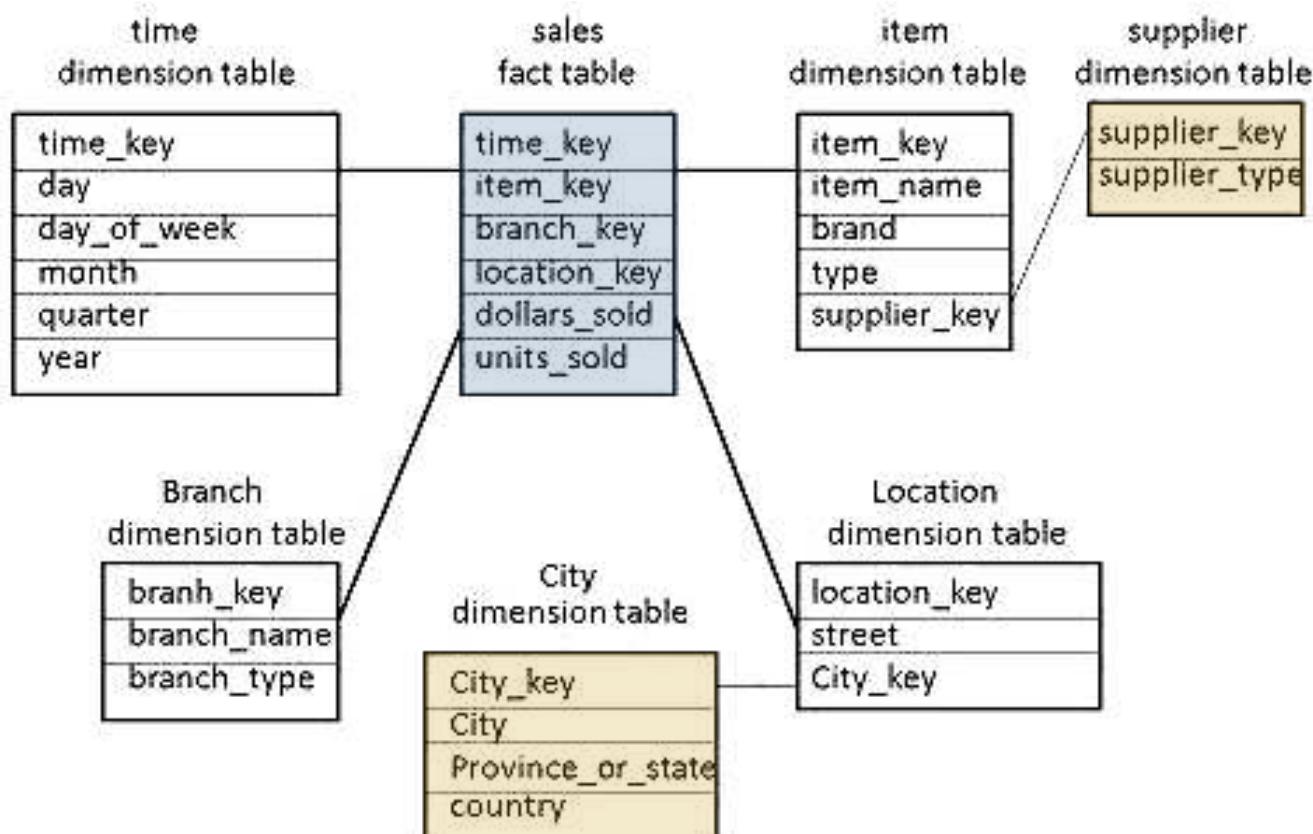
Приклад схеми “зірка”

- 1 факт: sales
- 4 виміри: time, branch, item, location
- 2 метрики: dollars_sold, units_sold



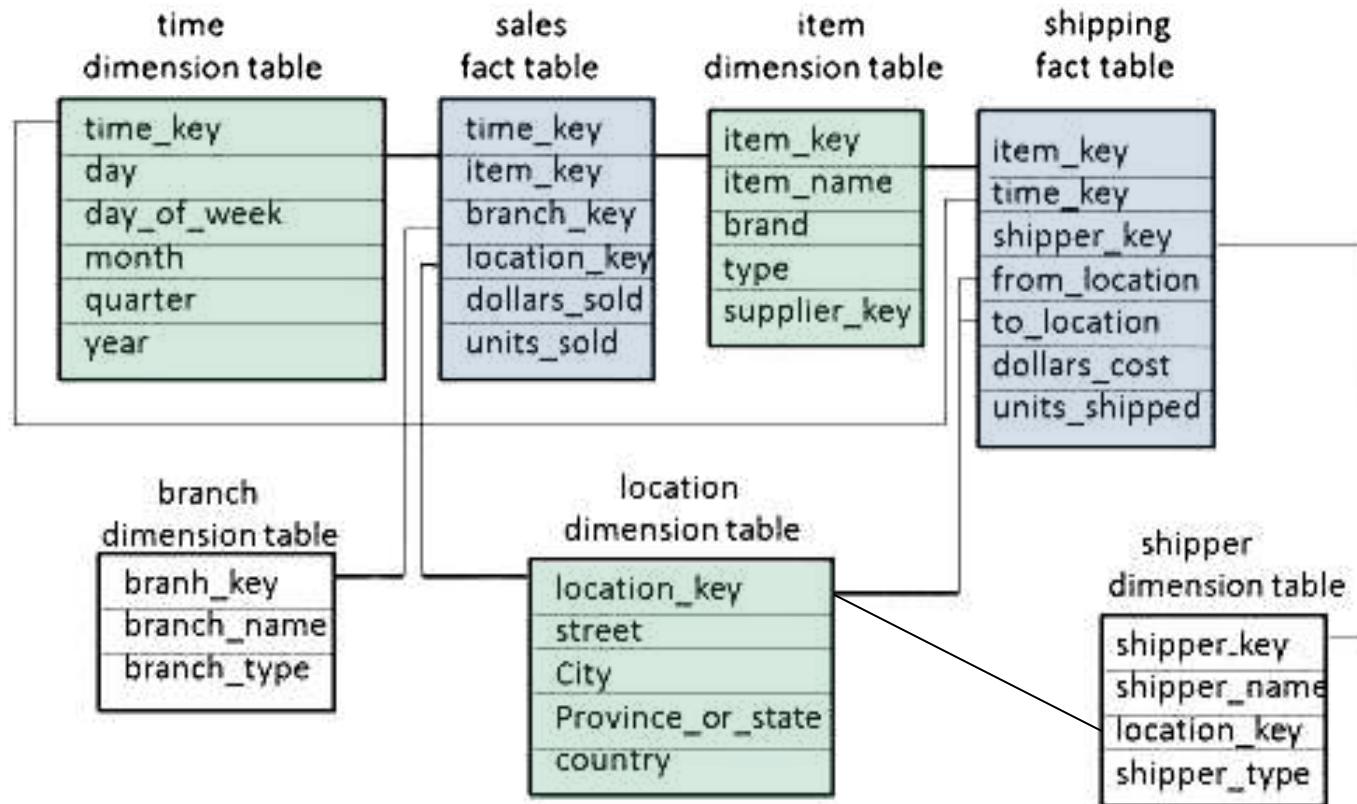
Приклад схеми “сніжинка”

- 1 факт: sales
- 4 виміри: time, branch, item + supplier, location + city
- 2 метрики: dollars_sold, units_sold

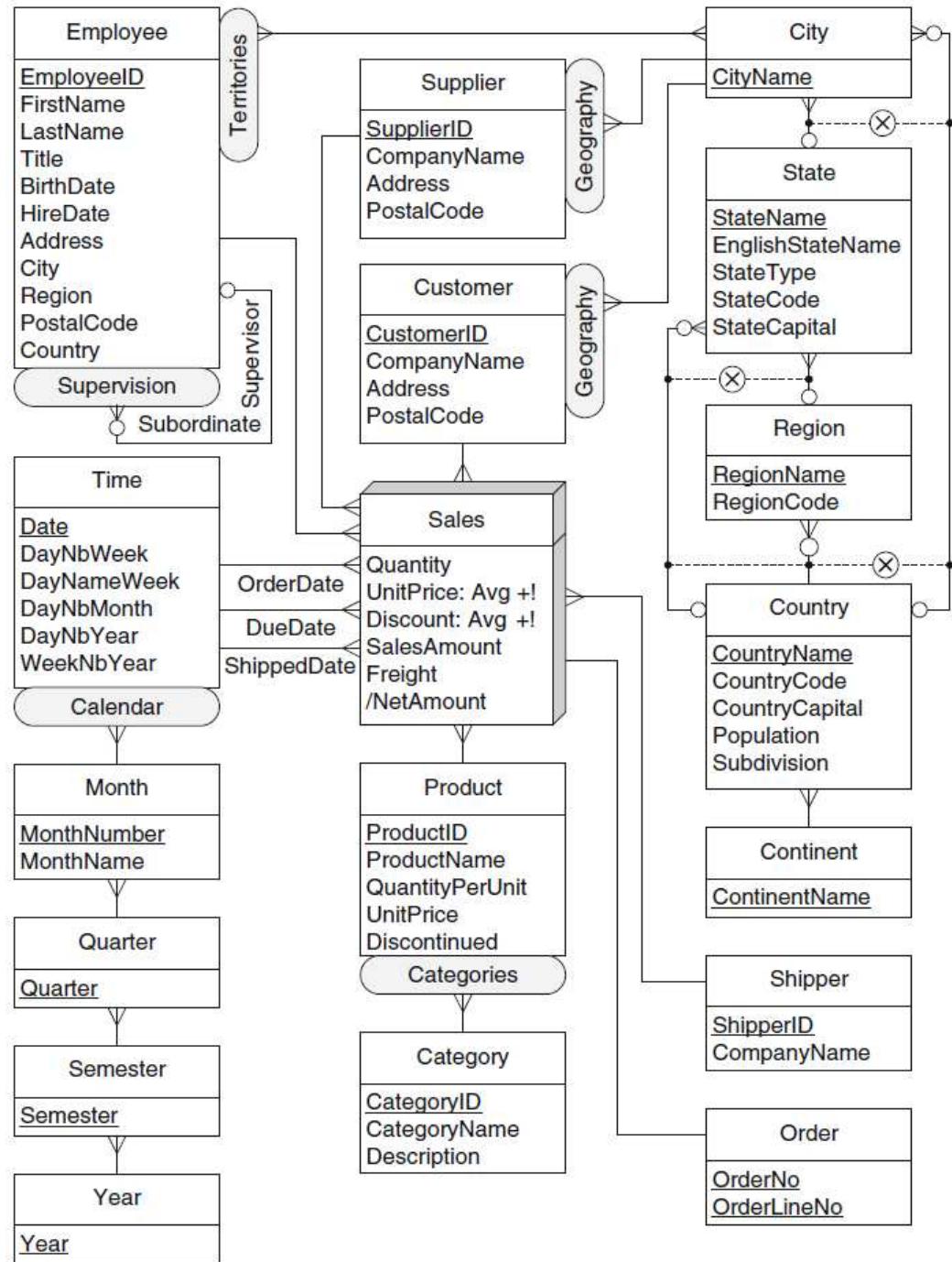


Приклад схеми “сузір’я”

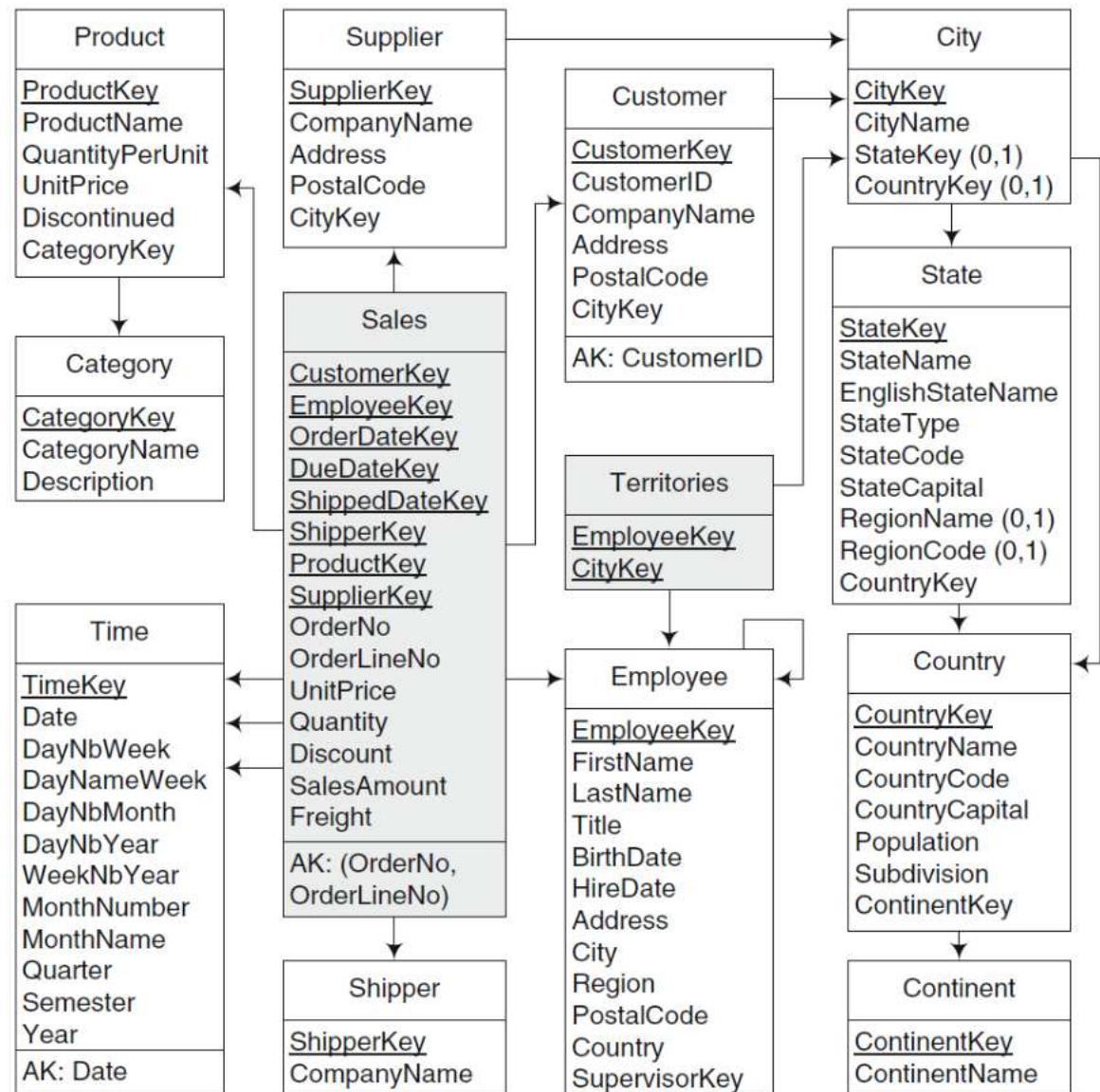
- 2 зв'язаних факти: sales, shipping
- 3 спільних виміри: item, time, location
- 4 метрики всього: dollars_sold, units_sold, dollars_cost, units_sh



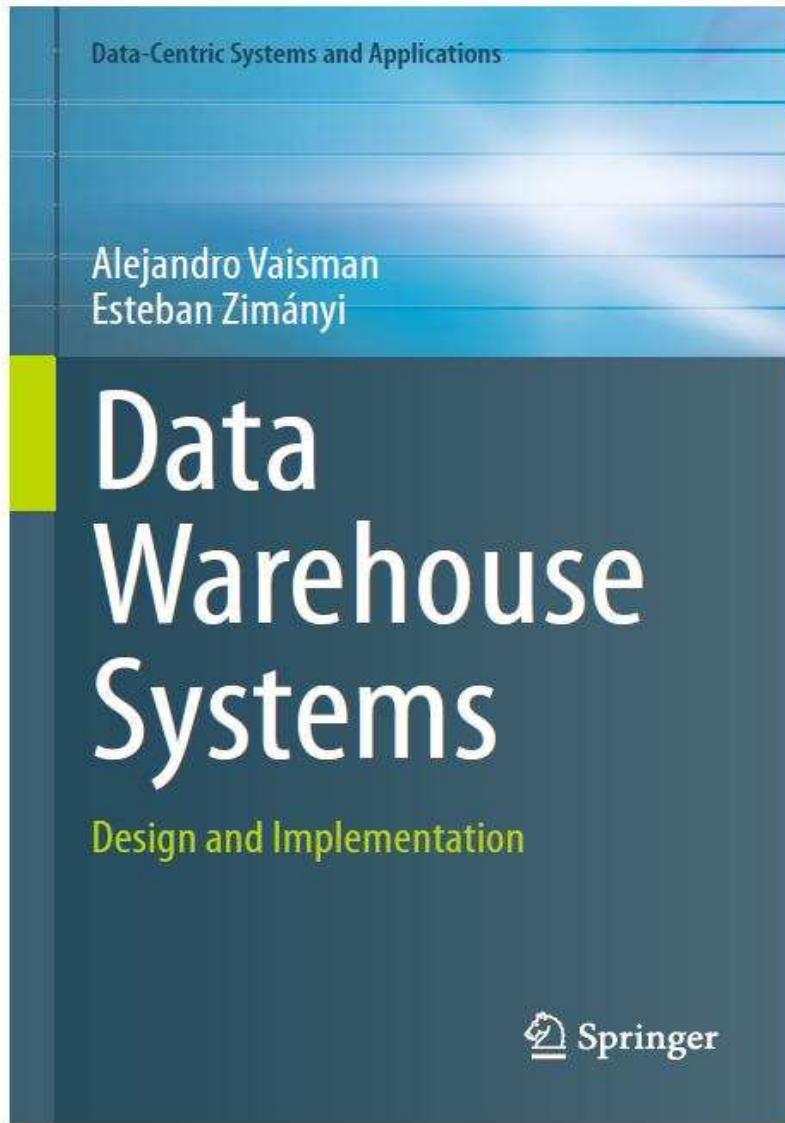
Conceptual Design of the Northwind warehouse:



Relational Representation of the Northwind warehouse:



Дякую за увагу!



The Data Warehouse Toolkit

Third Edition

The Definitive Guide
to Dimensional
Modeling

Ralph Kimball
Margy Ross



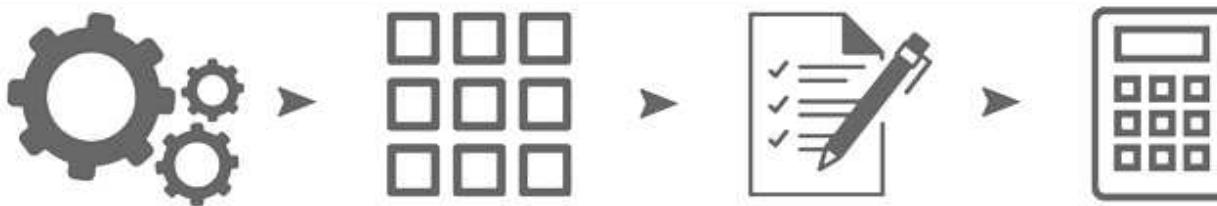
СХОВИЩА ДАНИХ: Лекція №5

НУ “Львівська Політехніка”, кафедра ПЗ

Проектування таблиць вимірів

Аналітика корпоративних даних

- Сутності предметної області
- Бізнес-процеси організації
- Процедури прийняття рішень
- Доступні корпоративні дані



1. Select the business process

3. Identify the dimensions

2. Declare the grain

4. Identify the facts

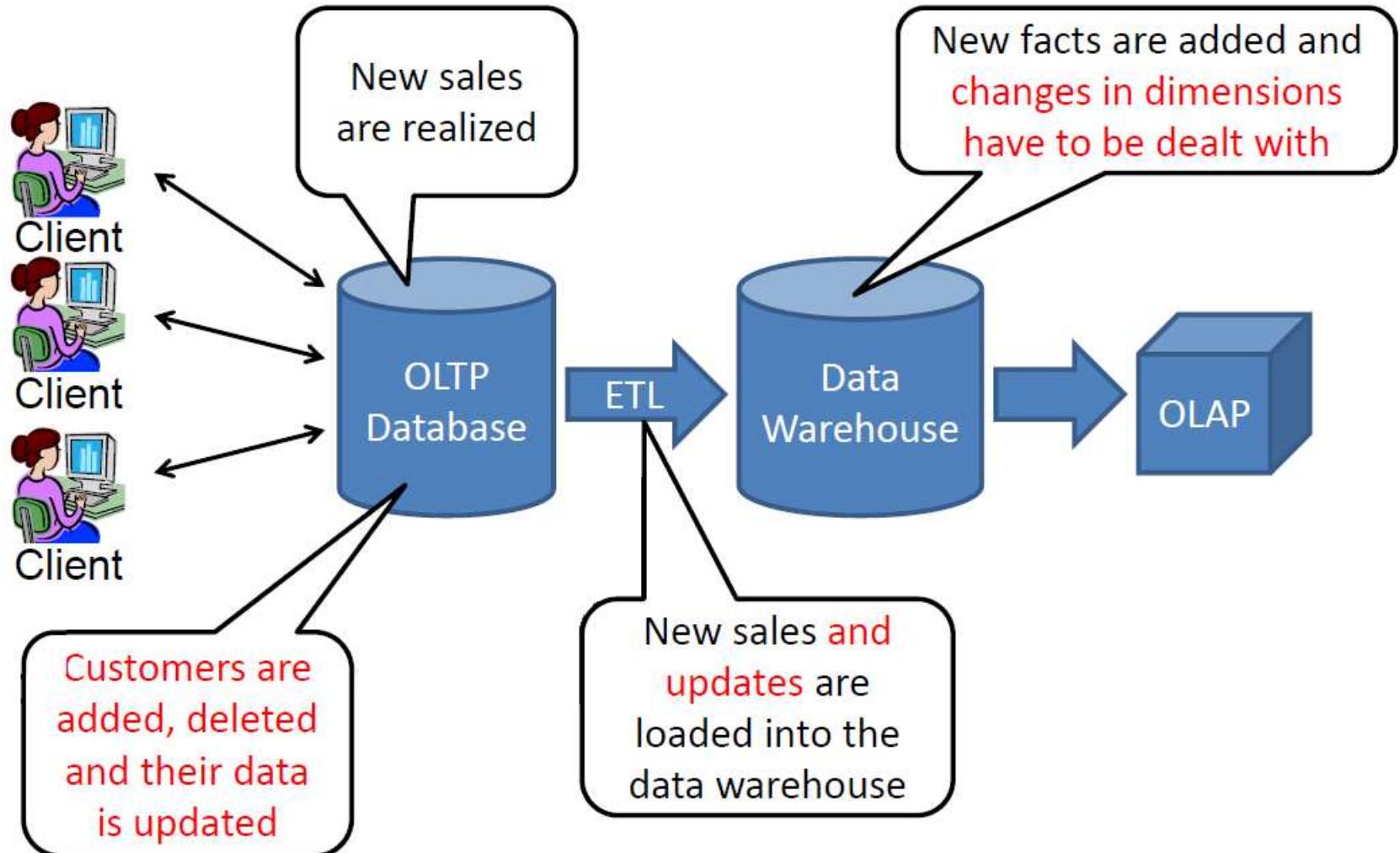
	Date	Customer	Product	Sales Rep	Deal	Warehouse	Shipper
Ordering	X	X	X	X	X		
Shipping to Customer	X	X	X		X	X	X
Receiving Payments	X	X		X			
Customer Returns	X	X	X	X	X	X	X

**Enterprise
Bus
Matrix**

Склад таблиці вимірів

- Сурогатний первинний ключ
 - Є зовнішнім ключем таблиці фактів
- Схема “зірка”, денормалізована
 - Назви елементів всіх рівнів виміру
 - Пусте значення назви дозволене
- Схема “сніжинка”, нормалізована
 - Назви елементів на одному рівні
 - Зовнішні ключі до інших таблиць ієархії
- Попередні значення
 - Для повільно змінного виміру
 - Проміжок часу актуальності значення
- Додаткові атрибути
 - Анотація, характеристики елементів
 - Супровідні дані для оновлення

Внесення нових даних – оновлення таблиць вимірів



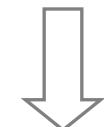
Нові дані виміру з часом (заміна значення – тип I повільний)

- Відслідковані нові дані з OLTP-таблиць переносяться шляхом оновлення відповідних колонок OLAP-таблиць вимірів

OLTP data
as of 2000

Customer		
CID	Name	Address
001	John	Dallas
002	Mary	Dallas
003	Pete	New York

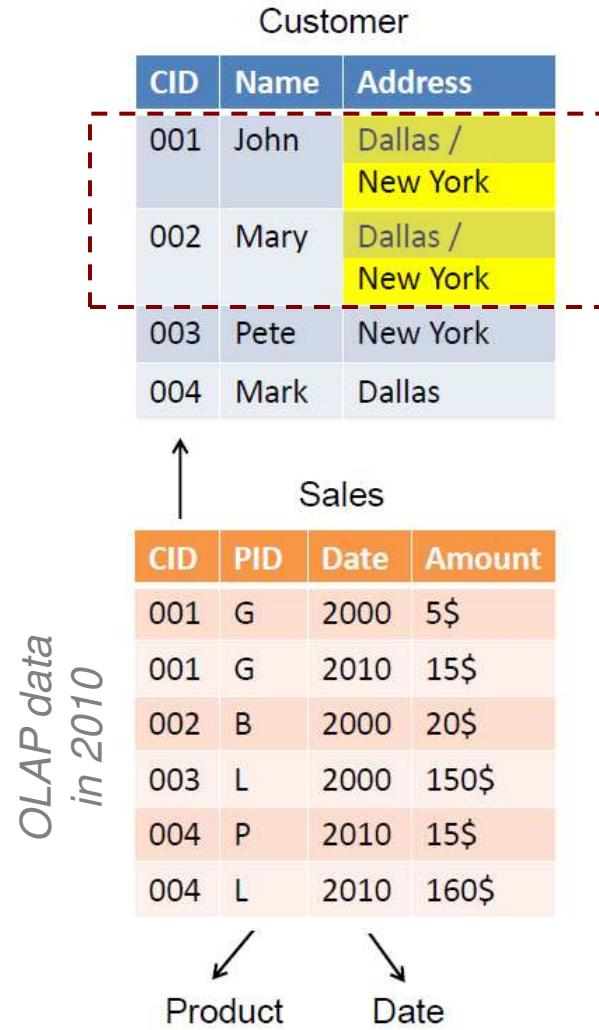
Sales		
CID	Product	Price
001	Gun	5\$
002	Beef	20\$
003	Lava lamp	150\$



Updates
in 2010

Customer		
CID	Name	Address
001	John	New York
002	Mary	New York
004	Mark	Dallas

Sales		
CID	Product	Price
001	Gun	15\$
004	Pork	15\$
004	Lava lamp	160\$



Варіативність значень виміру (новий рядок – тип II повільний)

- Створюється запис із новим сурогатним ключем, якій зберігає оновлені дані
- Фіксується час актуальності даних, що дозволяє відслідковувати зміни
- Додається колонка із маркером для позначення поточної версії елементів виміру

BookID	Book	Rating	Genre	ValidFrom	ValidTo	Newest
7493	Tropical Food	4 stars	Children's books	2006-03-01	2008-12-31	No
9436	Winnie the Pooh	5 stars	Children's books	2000-01-01	9999-12-31	Yes
9948	Gone With the Wind	4 stars	Fiction	1999-06-01	2008-10-15	No
9967	Italian Food	4 stars	Cooking	2003-04-05	2009-05-01	No
9995	Gone With the Wind	3 stars	Fiction	2008-10-16	9999-12-31	Yes
10100	Tropical Food	4 stars	Cooking	2009-01-01	9999-12-31	Yes
11319	Italian Food	4 stars	Mediterranean cooking	2009-05-02	9999-12-31	Yes

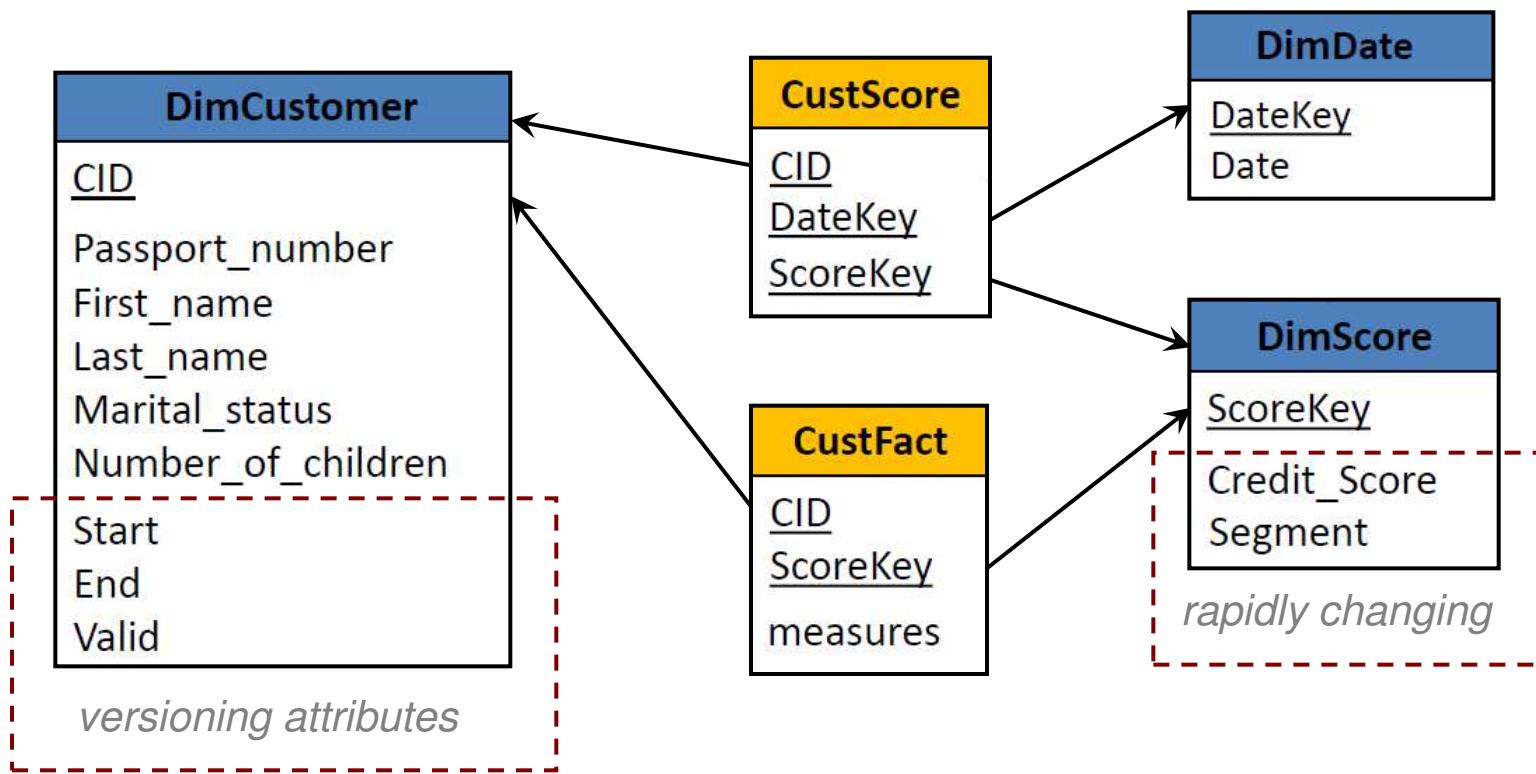
Варіативність значень виміру (додаткова колонка – тип III повільний)

- Характеристика виміру змінюється з часом, але первинне значення має залишатись для аналізу історичних даних
- Додається нова колонка для зберігання попереднього значення, рядок оновлюється
- Старих значень може бути декілька, тоді додаються колонки з нумерованими назвами
- За необхідністю фіксується дата останньої зміни

BookID	Book	Rating	OldRating	Genre	OldGenre
7493	Tropical Food	4 stars	4 stars	Cooking	Children's books
9436	Winnie the Pooh	5 stars	5 stars	Children's books	Children's books
9948	Gone With the Wind	3 stars	4 stars	Fiction	Fiction
9967	Italian Food	4 stars	4 stars	Mediterranean cooking	Cooking

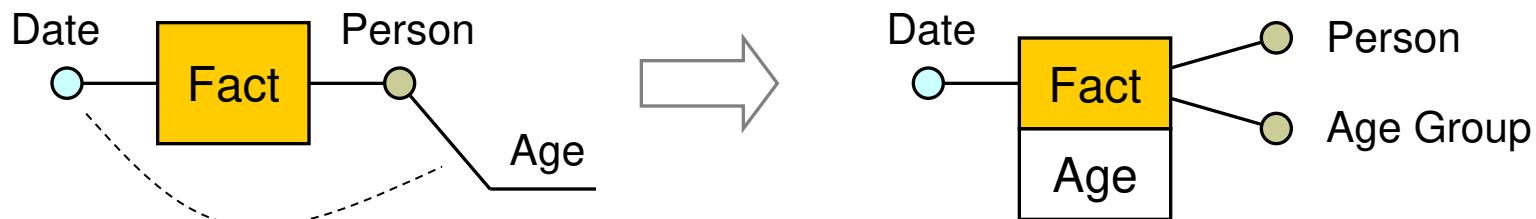
Варіативність значень виміру (декомпозиція – швидкозмінний тип)

- Перенесення із таблиці складеного виміру швидкозмінних характеристик в окрему таблицю
- Створення таблиці фактів без метрик, що з'єднує обидва виміри на певну дату



Дискретизація значень виміру (Discrete Values Dimension)

- Suppose frequently changing attributes have small domains
 - We could force this situation by discretizing some attributes with many values
- Populate the dimension with possible values
- What if we need to keep the changes in the dimensions?
 - Make a new fact with the attribute as measure



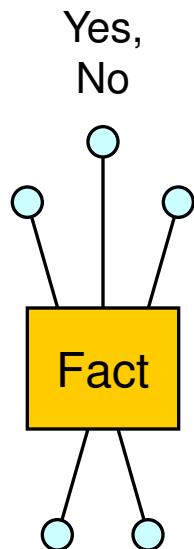
Вимір дати та часу (Date and Time Dimension)

- Date-time dimensions can become large
 - Enormous number of possible combinations
 - Either get from data (expensive) or generate all possibilities (infeasible)
- Therefore: usually split into Date dimension at granularity day and a Time-of-day dimension
 - Limited number of dates
 - Only 1440 minutes in a day



Вимір дрібних атрибутів (Junk Dimension)

- Many small dimensions combined
 - Low-cardinality indicators
 - Typically flags; promotion; how-displayed
- Combine into one dimension, fully populate

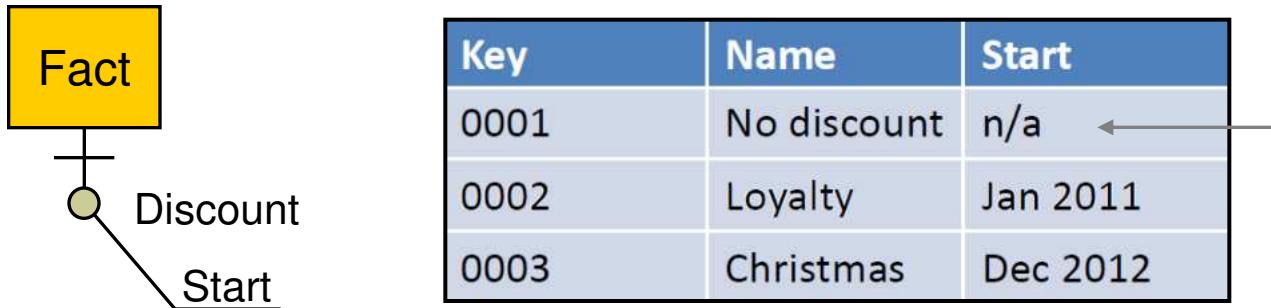


JunkID	Packed	Shipped	Delivered	Returned	Refunded
001	N	N	N	N	N
002	N	N	N	N	Y
003	N	N	N	Y	N
004	N	N	N	Y	Y
...
032	Y	Y	Y	Y	Y

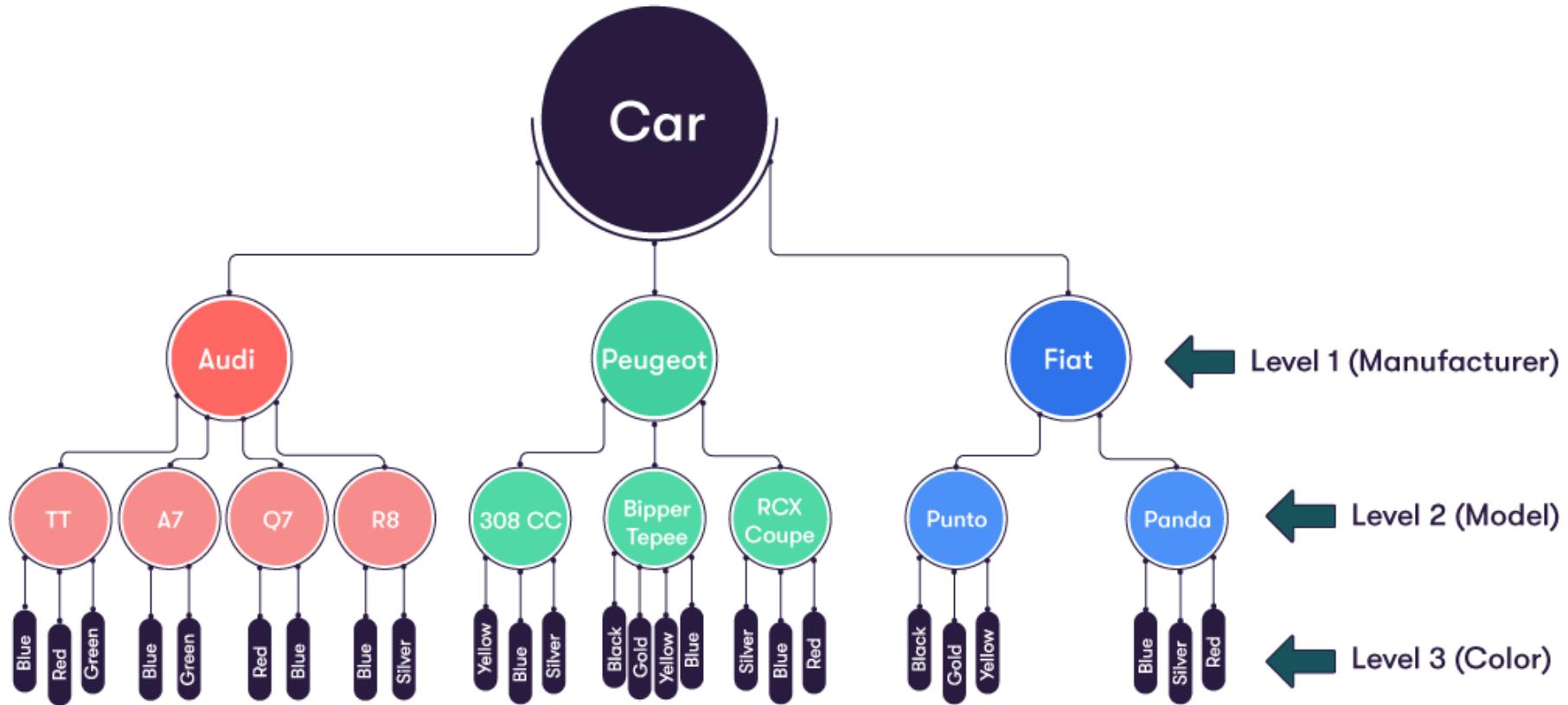
Cartesian Product

Необов'язковий вимір (Optional Dimension)

- Avoid the use of “null”
 - Confusing and non-descriptive
 - Special case for queries ($\text{null} \neq \text{null}$)

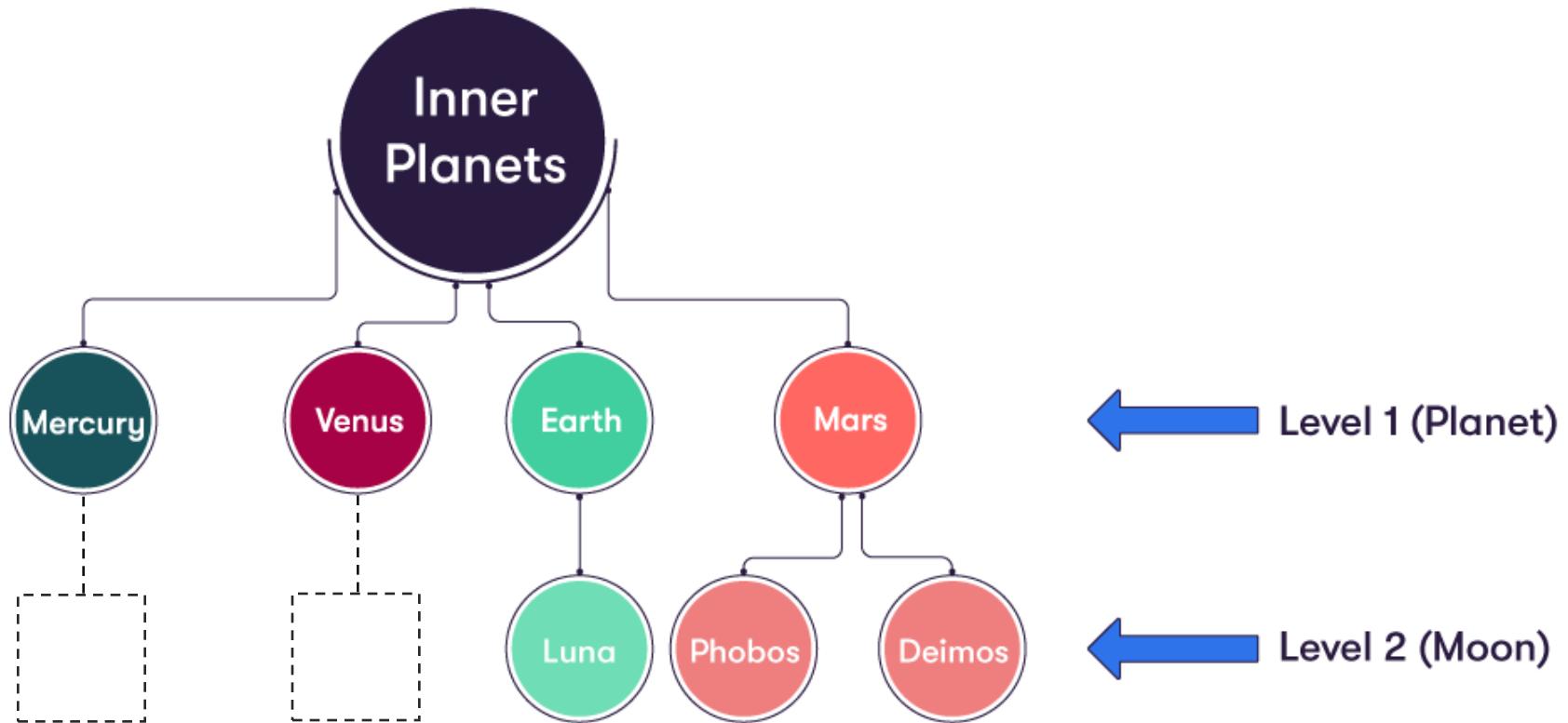


Збалансована ієрархія виміру (Symmetrical Hierarchy)



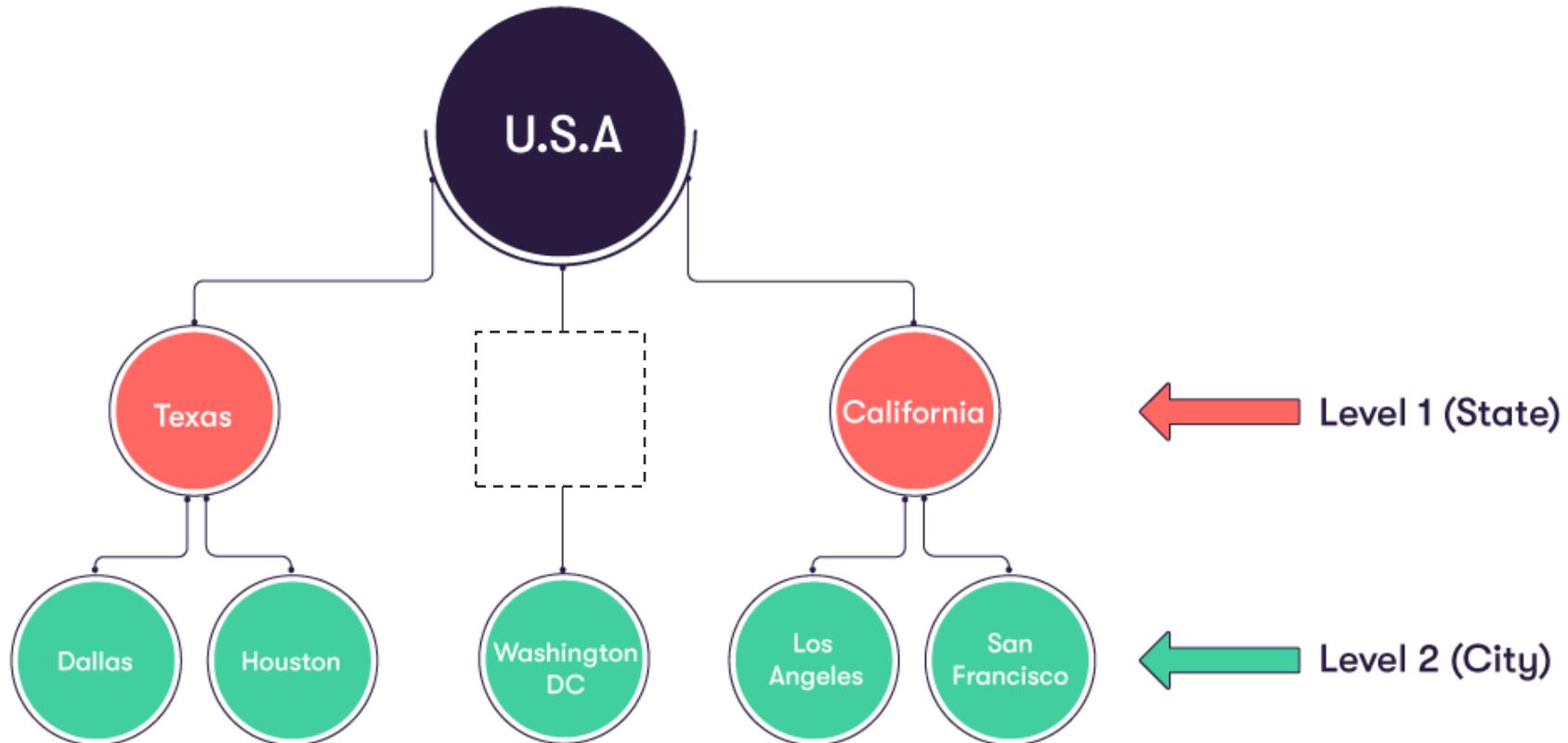
- Наперед відома кількість та назви рівнів
- Відношення між сущностями один до багатьох
- Рівні є колонкам в денормалізованій таблиці

Незбалансована ієрархія виміру (Asymmetrical Hierarchy)



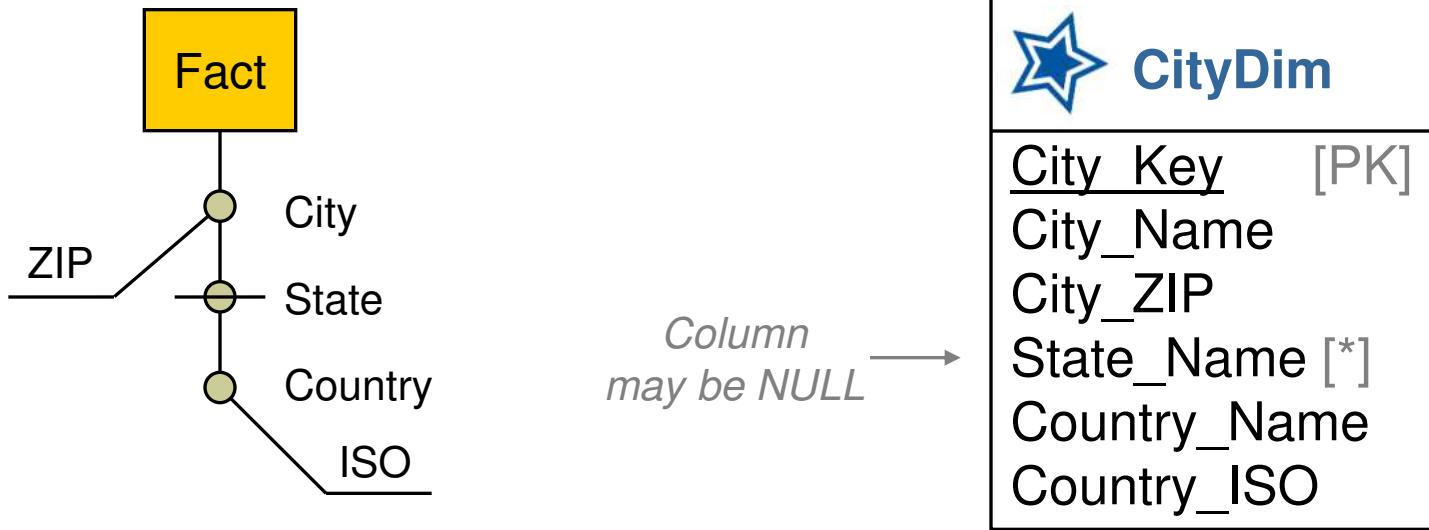
- Обмежена кількість рівнів без спільних елементів
- Допускається відсутність даних на кінцевих рівнях
- Представлення таблицею із NULL-колонками

Нерівна ієрархія виміру (Ragged Hierarchy)



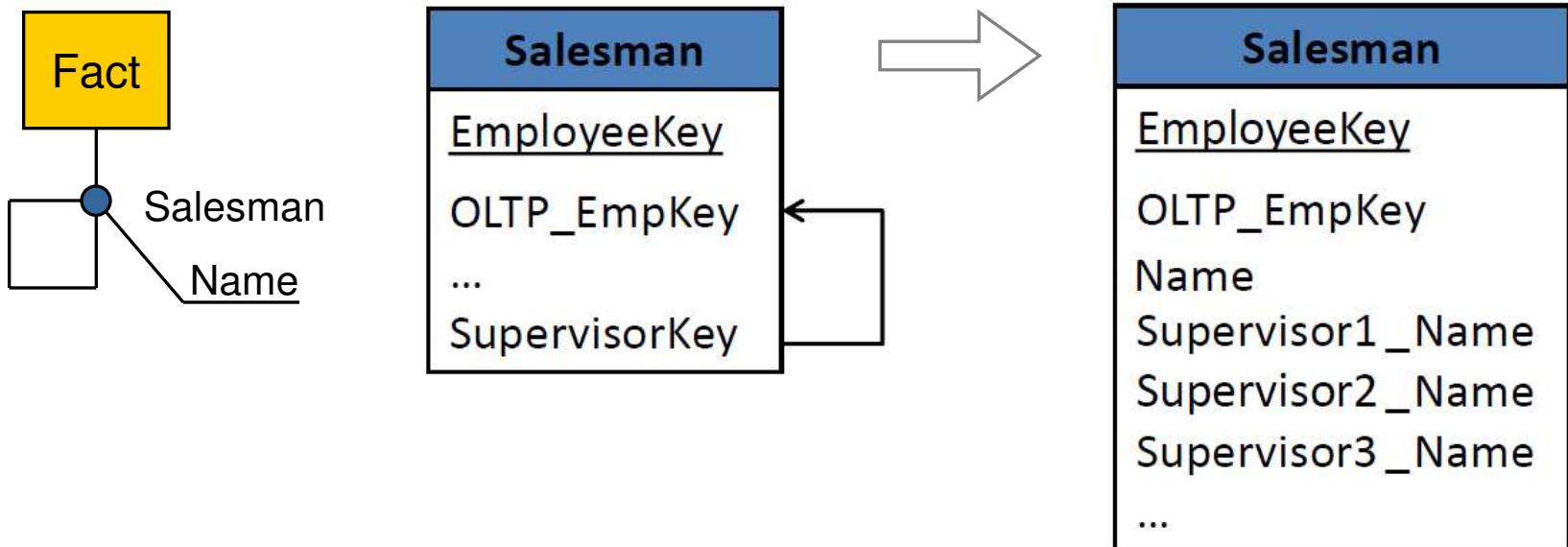
- Невелика кількість рівнів із можливим пропуском
- Виконується заміна NULL-значення при агрегації
- Або балансування переносом даних з нижнього рівня

Приклад реалізації таблиць виміру схемою “зірка” та “сніжинка”



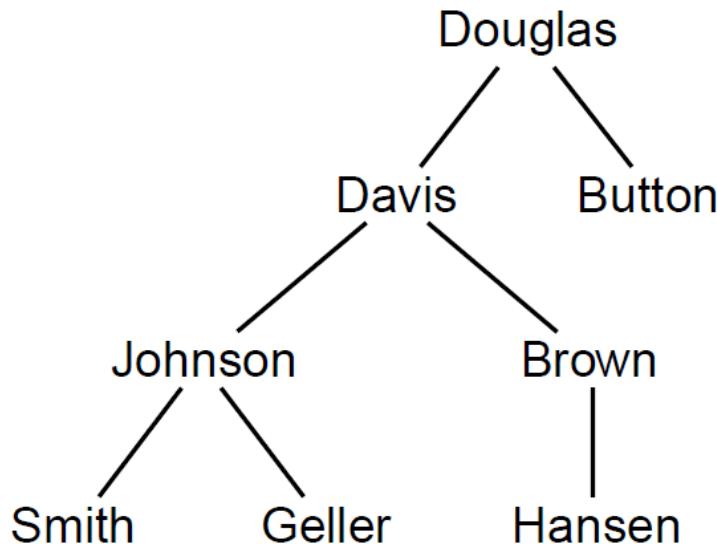
Обмежена рекурсивна ієрархія (Fixed-Levels Recursive Hierarchy)

- Розкриття зв'язків в одну денормалізовану таблицю для обмеженої глибини дерева
- Використання напрямку обходу “parent”, що забезпечує унікальність комбінації полів
- Назва колонки відповідає рівню вкладеності



Довільна рекурсивна ієархія (Parent-Child Strict Hierarchy)

- Відношення один до багатьох в ієархії
- Довільна глибина вкладеності елементів
- Потребує створення нормалізованої таблиці-мосту, яка містить всі шляхи від кореня до кожного вузла із фіксацією відстані



<u>EmployeeID</u>	<u>Employee</u>	<u>ManagerID</u>
1	Douglas	NULL
2	Davis	1
3	Johnson	2
4	Smith	3
5	Geller	3
6	Brown	2
7	Hansen	6
8	Button	1

Таблиця-міст виконує роль навігатора по ієрархії виміру куба

Ancestor: A foreign key column referencing the primary key column of the dimension table to capture a row's ancestor

Descendant: A foreign key column referencing the primary key column of the dimension table to capture a row's descendant

Distance: An integer capturing the length of the path between *Ancestor* and *Descendant*

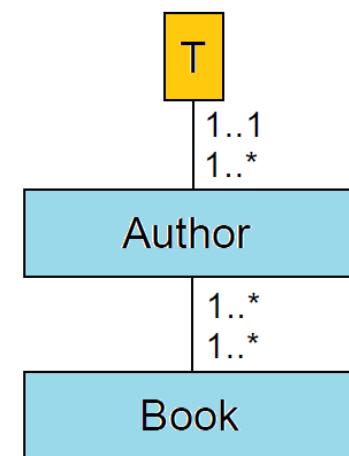
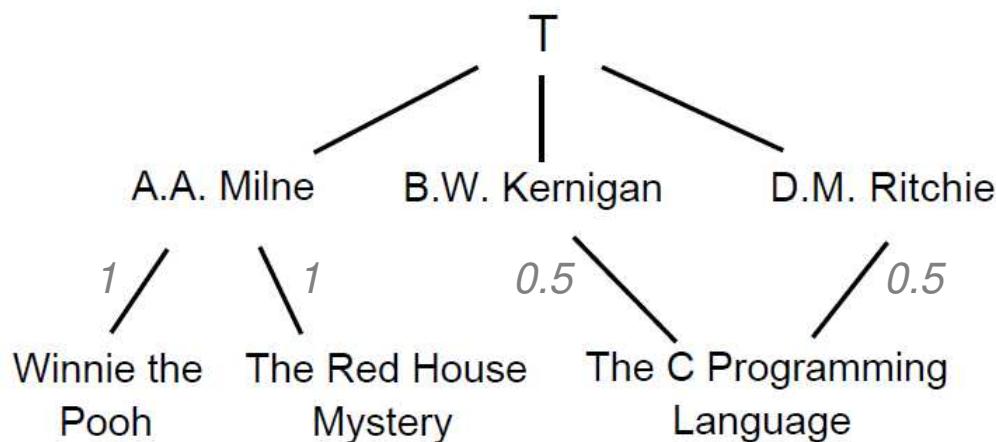
Bottom Flag: A Boolean value indicating whether *Descendant* is at the lowest level (i.e., does not have descendants)

Top Flag: A Boolean value indicating whether *Ancestor* is at the top-most level (i.e., does not have ancestors)

Ancestor	Descendant	Distance	Bottom Flag	Top Flag
1	1	0	False	True
1	2	1	False	False
1	8	1	True	False
1	3	2	False	False
1	6	2	False	False
1	4	3	True	False
1	5	3	True	False
1	7	3	True	False
2	2	0	False	False
2	3	1	False	False
2	6	1	False	False
2	4	2	True	False
2	5	2	True	False
2	7	2	True	False
8	8	0	True	False
3	3	0	False	False
3	4	1	True	False
3	5	1	True	False
6	6	0	False	False
6	7	1	True	False
4	4	0	True	False
5	5	0	True	False
7	7	0	True	False

Множинність зв'язків в ієрархії (Many-to-Many Non-Strict Hierarchy)

- Відношення багато до багатьох між рівнями
- При агрегації фактів є вплив на значення метрики, який приводить до повторного врахування
- Таблиця-міст має ваговий множник для кожної гілки відношення, щоб розділити адитивні факти
- Сума гілок на елемент нижчого рівня складає 100%



Таблиця-міст забезпечує операцію розгортки куба на складові, зберігаючи суму

Percentage shows how large a fraction is “owned by” a parent

Order is just a positive integer to keep a position in the list

When loading data into the database, it is important that these property values are assigned correctly such that percentages add up to 1.0

<u>BookID</u>	<u>AuthorID</u>	Order	Percentage
1	1	1	1.0
2	2	1	0.5
2	3	2	0.5
3	1	1	1.0

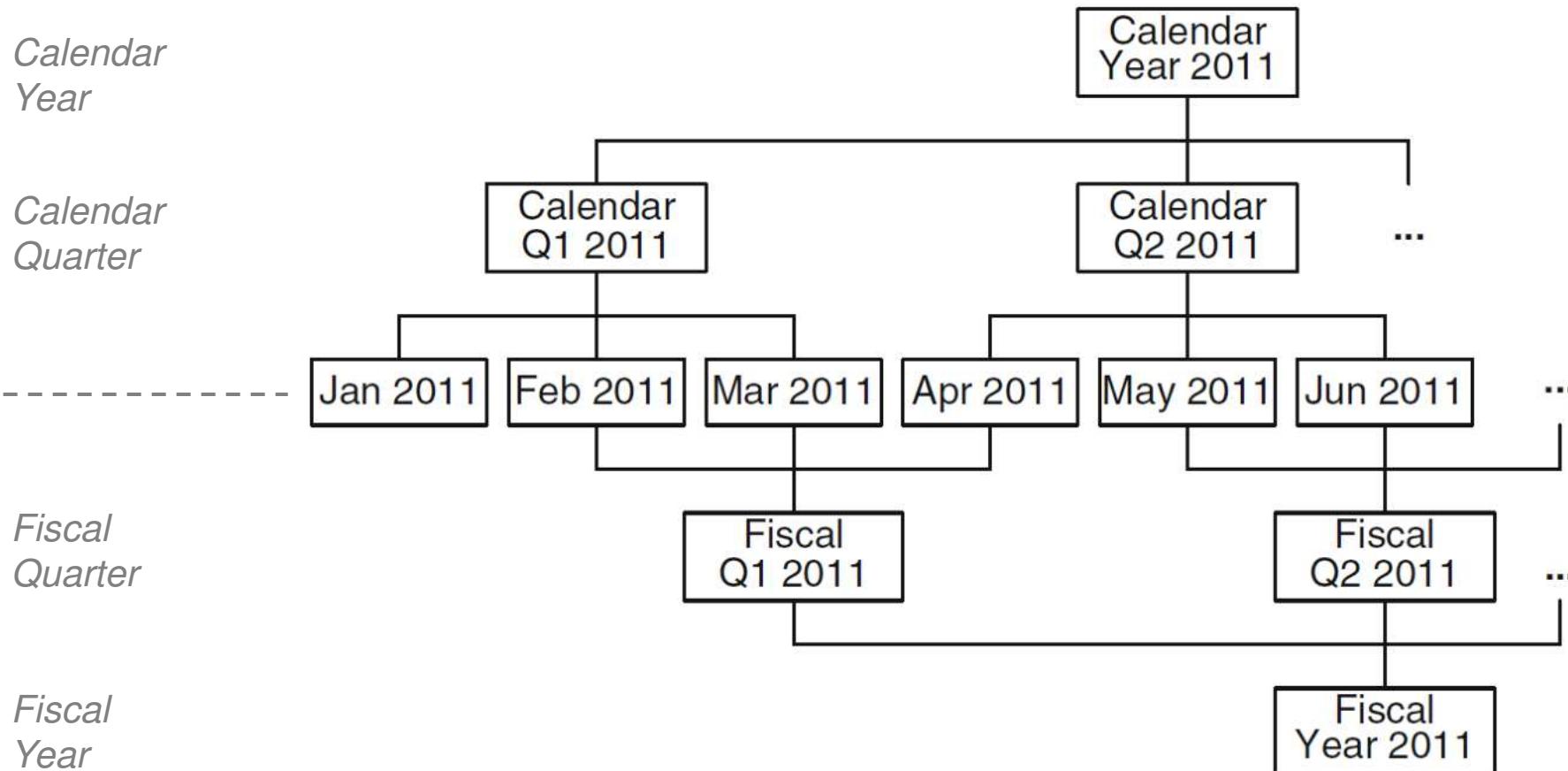
<u>BookID</u>	Title
1	Winnie the Pooh
2	The C Programming Language
3	The Red House Mystery

<u>AuthorID</u>	Name	...
1	A.A. Milne	...
2	B.W. Kernigan	...
3	D.M. Ritchie	...

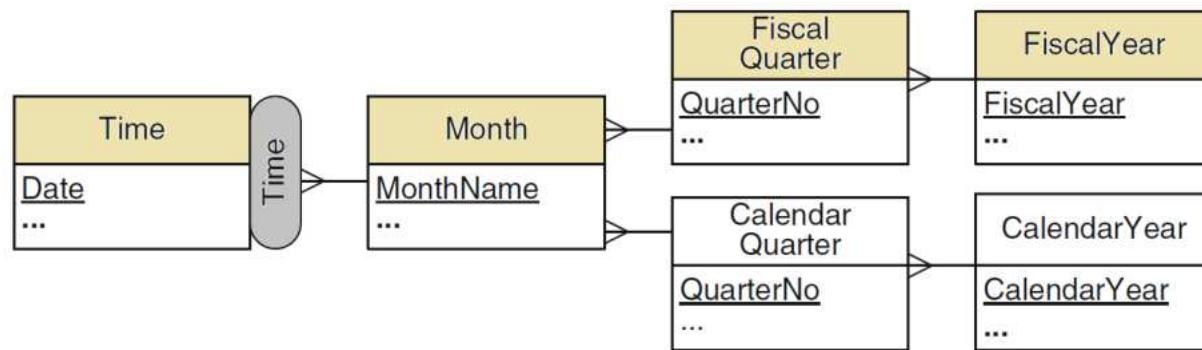
- Якщо зв'язок багато до багатьох є між фактом та виміром, то слід створити таблицю “фактів без метрик”

Паралельні гілки в ієрархії (Parallel Hierarchy)

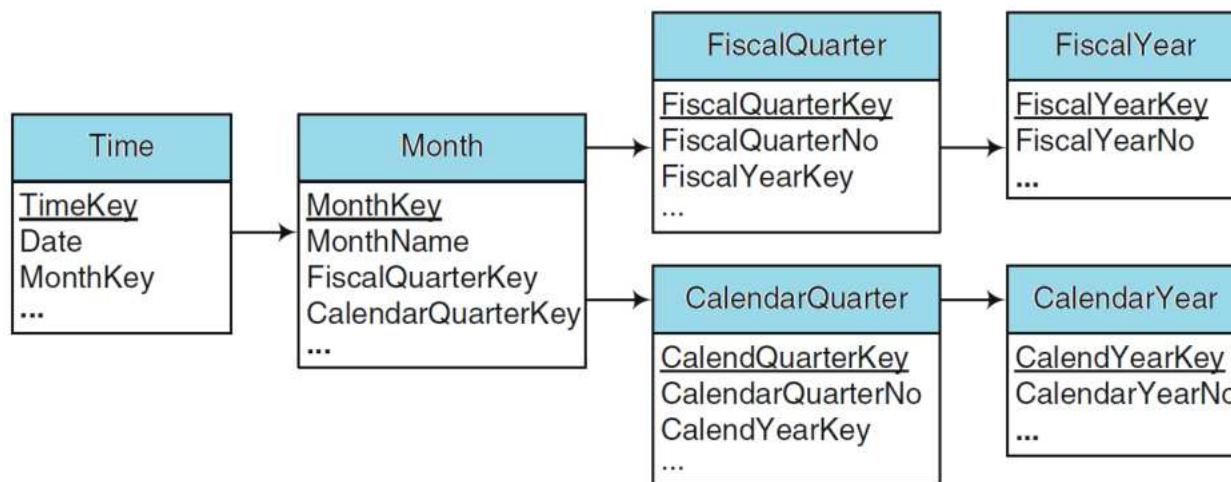
- Вимір надає різні шляхи для аналітичних завдань
- Ієрархія має спільний початковий вузол, відносно якого побудовано зв'язок один до багатьох



➤ Моделювання концептуальної схеми:

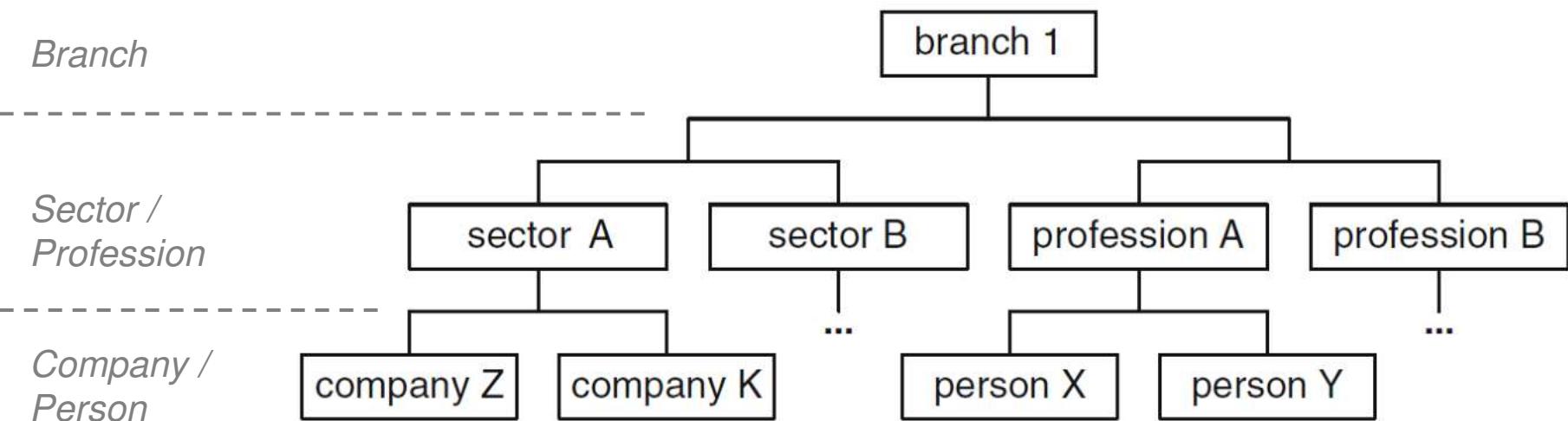


➤ Моделювання логічної схеми:

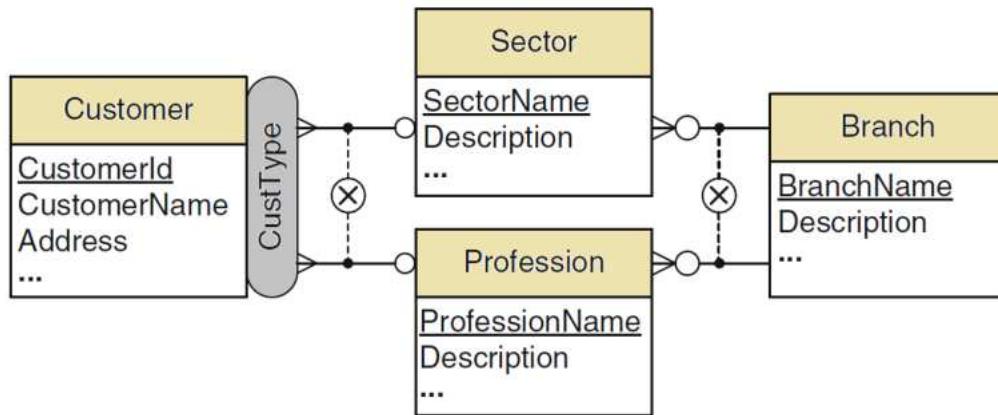


Генералізація із спільним рівнем (Shared-Level Generalized Hierarchy)

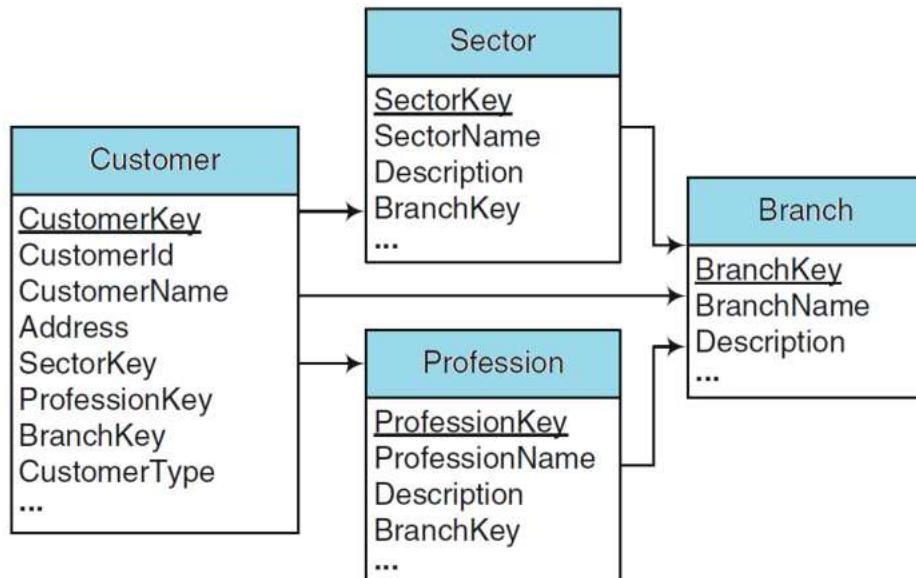
- Вимір розширяє звичайну ієрархію наслідування спільним кінцевим рівнем
- Аналітика по всіх значеннях можлива завдяки додаванню зв'язку на кінцевий рівень у базову таблицю виміру
- Потрібна підстановка значення-замінника в запитах агрегації, якщо елемент проміжного рівня не відповідає гілці ієрархії



➤ Моделювання концептуальної схеми:



➤ Моделювання логічної схеми:

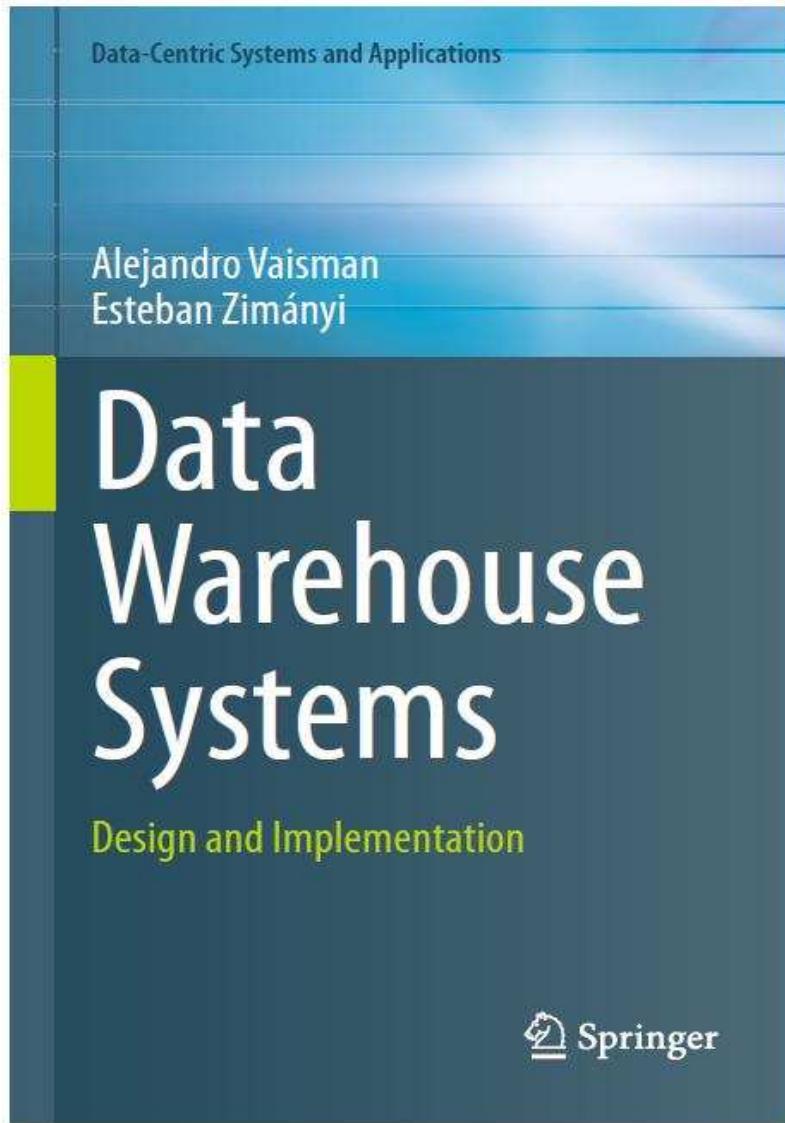


```
ALTER TABLE Customer ADD CHECK  
(CustomerType IN  
('Person', 'Company'))
```

```
ALTER TABLE Customer ADD CHECK  
((CustomerType != 'Person') OR  
(ProfessionKey IS NOT NULL  
AND SectorKey IS NULL))
```

```
ALTER TABLE Customer ADD CHECK  
((CustomerType != 'Company') OR  
(ProfessionKey IS NULL  
AND SectorKey IS NOT NULL))
```

Дякую за увагу!



The Data Warehouse Toolkit

Third Edition

The Definitive Guide
to Dimensional
Modeling

Ralph Kimball
Margy Ross



СХОВИЩА ДАНИХ: Лекція №6

НУ “Львівська Політехніка”, кафедра ПЗ

Проектування таблиць фактів

Ідентифікація метрик та вимірів

- **Ступінь гранулювання**
 - Які показники об'єднувати на якому рівні?
- **Відповідність вимірам**
 - Які властивості сутностей ідентифікують факти?
- **Поседнання процесів**
 - Як консолідувати показники різних бізнес-процесів?
- **Попереднє обчислення**
 - Які значення походять від декількох характеристик?
- **Завершеність подій**
 - Що саме визначає поточний стан або результат?
- **Актуальність для аналізу**
 - Чи буде властивість корисною для прийняття рішень?
- **Трансляція об'єктів**
 - Які назви будуть зручними для бізнес-аналітиків?
- **Трансформація зв'язків**
 - Якими будуть виміри з урахуванням можливих змін первинних даних?
- **Часові атрибути**
 - Яким чином фіксувати час актуальності даних?
- **Архівні або проміжні дані**
 - Чи потрібно акумулювати дані, що є тимчасовими в оперативній базі?
- **Прогнозована стабільність**
 - Коли значення можуть змінюватись у сховищі?
- **Регулярність оновлення**
 - Якою буде періодичність завантаження даних?

Склад таблиці фактів

- Зовнішні ключі до таблиць вимірів
 - Пусте посилання заборонено
- Прості значення вироджених вимірів
 - Частина первинного ключа
- Колонки числових метрик
 - Пусте значення дозволене
- Сурогатний альтернативний ключ
 - Для зручності ідентифікації рядка
- Довідкові супровідні колонки
 - Для зовнішніх систем та оновлення
- Індекс на вимір дати
 - Пришвидшує операцію часового зрізу



Приклад: аналітика фінансових операцій

Властивість	Значення	Зразок
Дата, час	До хвилини	2025.03.17 14:50
Тип	Плюс, мінус	Кредит
Призначення	Категорія	Покупка
Рахунок	Назва	Основний
Сума	грн	500
Податків нараховано	грн	3,6
Комісії нараховано	грн	2,5

Транзакційна таблиця фактів

Дата	Рахунок	Операція	Сума	Час
1 березня	Основний	Зарплата	+200	10:00
1 березня	Основний	Покупка	-20	12:00
2 березня	Основний	Покупка	-100	14:00
2 березня	Бонусний	Кешбек	+10	14:00
3 березня	Основний	Покупка	-50	16:00
3 березня	Бонусний	Покупка	-5	16:00
3 березня	Основний	Покупка	-30	18:00
4 березня	Основний	Поповнення	+100	10:00

об'єднані джерела даних

A row in a transaction fact table corresponds to a measurement event at a point in space and time. Atomic transaction grain fact tables are the most dimensional and expressive fact tables.

вимір із гілками:
- призначення
- дебет/кредит

адитивна метрика

вироджений вимір

Таблиця періодичних знімків

Дата	Рахунок	Дебет	Кредит	Дельта часу
1 березня	Основний	200	20	2
2 березня	Основний		100	0
2 березня	Бонусний	10		0
3 березня	Основний		80	2
3 березня	Бонусний		5	0
4 березня	Основний	100		0

↑
інтервал часу
(повна доба)

↑
адитивні метрики
на основі категорії

розраховане значення

A row in a periodic snapshot fact table summarizes many measurement events occurring over a standard period, such as a day, a week, or a month. The grain is the period, not the individual transaction.

неадитивна міра
інтенсивності (сума за
вимірами не має сенсу)

Таблиця кумулятивних знімків

Дата	Рахунок	Залишок	Операцій
1 березня	Основний	180	2
1 березня	Бонусний	0	0
2 березня	Основний	80	1
2 березня	Бонусний	10	1
3 березня	Основний	0	2
3 березня	Бонусний	5	1
4 березня	Основний	100	1
4 березня	Бонусний	5	0

Значення агрегатної функції
на кінець встановленої дати

A row in an accumulating snapshot fact table summarizes the measurement events occurring at predictable steps between the beginning and the end of a process. Pipeline or workflow processes, that have a defined start point, standard intermediate steps, and defined end point can be modeled.

↑
адитивна метрика
напівадитивна метрика
(неможливе додавання за часом)

Таблиця заголовків фактів

Дата	Рахунок	Операція	Податків	Комісій	Додаткові дані
1 березня	Основний	Зарплата	36	1	
1 березня	Основний	Покупка	4		
2 березня	Основний	Покупка	20		
2 березня	Бонусний	Кешбек	1,8		
3 березня	Основний	Покупка	16		
3 березня	Бонусний	Покупка	1	0,5	
4 березня	Основний	Поповнення		2	

Operational systems often consist of a transaction header row that's associated with multiple transaction lines. With these schemas, all the header-level dimension foreign keys and degenerate dimensions should be included on the line-level fact table.

метрики характеризують групи фактів за вимірами транзакційної таблиці (зavedені показники або зв'язки з іншими вимірами)

Таблиця агрегатів фактів

Дата	Рахунок	Операція	Сума	Кількість
березень	Основний	Зарплата	+200	1
березень	Основний	Поповнення	+100	1
березень	Основний	Покупка	-200	4
березень	Бонусний	Кешбек	+10	1
березень	Бонусний	Покупка	-5	1
березень	Основний	всі	+100	6
березень	Бонусний	всі	+5	2
березень	всі	Дебет	+310	3
березень	всі	Кредит	-205	5

Aggregate fact tables are simple numeric rollups of atomic fact table data built solely to accelerate query performance. These tables should be available to the reporting layer at the same time as the atomic fact tables so that tools choose the appropriate aggregate level.

↑
обрані агрегатні функції метрик
↑
обраний рівень виміру

агрегація значень по вимірах



Таблиця фактів: можливості аналізу

- Один процес – декілька таблиць
 - Дії, підсумки, зв'язки, консолідація, обчислення
- Додаткові метрики
 - Міри інтенсивності, поточні характеристики
- Додаткові виміри
 - Дискретизація, дані різних рівнів, бізнес-ключі

Визначення додаткових метрик

- **Консолідована таблиця фактів** – комбінація метрик різних бізнес-процесів для аналізу, що відповідають рівню деталізації таблиці
- **Метрика як інтервал часу між подіями** – обсяг пройденого часу від початку або від попередній стадії бізнес-процесу

Створено
Прийнято
В дорозі
Прибуло
Отримано

Зведені показники
<u>Місяць</u>
<u>Магазин</u>
Кількість чеків
Продажі, грн (б.п. №1)
Витрати, грн (б.п. №2)
Прогноз, грн (б.п. №3)

Трекінг посилок
<u>Дата</u>
<u>Ідентифікатор</u>
<u>Відправник</u>
<u>Отримувач</u>
<u>Попередній стан</u>
<u>Поточний стан</u>
Час від створення, год
Час від зміни стану, год

Визначення додаткових вимірів

- **Числові характеристики вимірів** – додавання у таблицю фактів метрики із створенням виміру по діапазону її значення

атрибути
продукту {

Аналітика пакування
<u>Дата</u>
<u>Замовник</u>
<u>Продукт</u>
<u>Вага</u> (<1, 1–3, 3–5, >5 кг)
<u>Об'єм</u> (мал, серед, великий)
Вага, грами
Об'єм, см куб
Вартість, грн

- **Факт на різних рівнях грануляції** – додавання необов'язкового зовн. ключа на рядок таблиці вимірів відповідного рівня ієархії



Нарахування плати
<u>Дзвінок</u> (момент)
<u>Дата</u> (щоденно)
<u>Місяць</u> (абоплата)
<u>Абонент</u>
<u>Тариф</u>
<u>Напрямок</u>
Нараховано, грн
Тривалість, хв

} або

Проблема багатозначних вимірів

- Зв'язок **M-M** між фактом та виміром дублює рядок із значенням метрики
- Повторне обчислення при агрегації даних
- Вирішення зміною структури аналітичних вимірів бізнес-процесу

$300 = \Sigma$

<u>Time</u>	<u>Account</u>	<u>Client</u>	Balance
T1	A1	C1	100
T1	A1	C2	100
T1	A1	C3	100
T1	A2	C1	500
T1	A2	C2	500

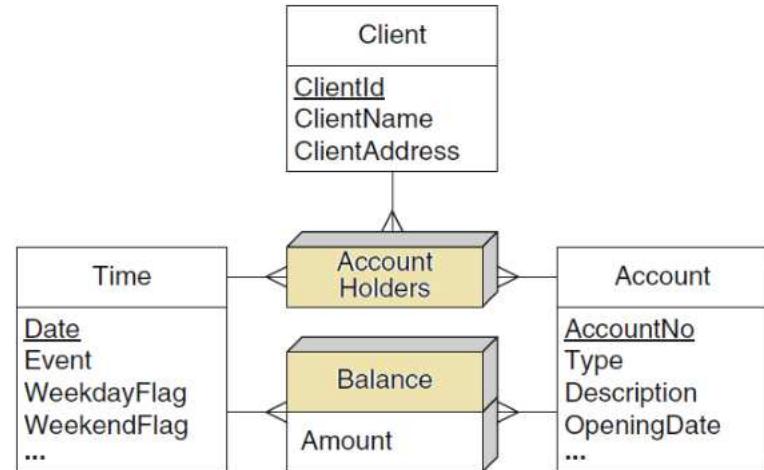
банківські рахунки
є спільними для
декількох клієнтів,
метрика балансу
має залишатись
напівадитивною

Врахування багатозначних вимірів

1

Створення факту без метрик, що окремо зв'язує відповідні виміри.

Перенесення зв'язку **1-М** до нього. В запитах слід поєднувати обидва факти.

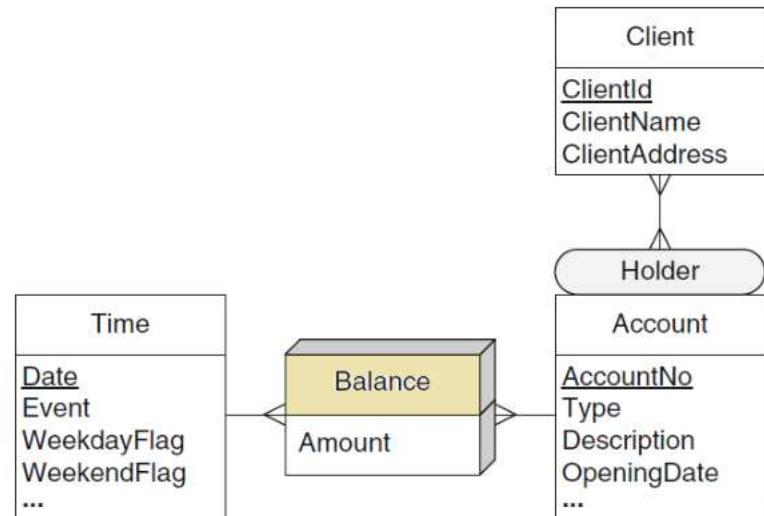


АБО

2

Перенесення відношення **М-М** в ієрархію існуючого виміру. Або створення штучного виміру-групи.

Таблиця-міст має бути застосована для зв'язування з використанням ваги чи атрибуту ексклюзивності.



Денормалізація для оптимізації

- **Фіксовані значення виміру першого рівня формують масив метрик** – зменшує кількість рядків у таблиці фактів, але потрібна окрема процедура агрегації по рівнях виміру

Аналітика обсягу продажів														
<u>Рік</u>	<u>Магазин</u>	<u>Товар</u>	січ.	лют.	бер.	кві.	тра.	чер.	лип.	сер.	вер.	жов.	лист.	гру.
рівень <u>“рік”</u>														метрика таблиці періодичних знімків по рівню <u>“місяць”</u>

- **Зовнішні ключі до нормалізованих таблиць вимірів** – спрощує агрегацію по рівнях ієрархії, але збільшує обсяг таблиці фактів

Аналітика обсягу продажів								
<u>День</u>	<u>Місяць</u>	<u>Рік</u>	<u>Магазин</u>	<u>Місто</u>	<u>Товар</u>	<u>Група</u>	<u>Бренд</u>	Сума, грн
↓ Дати –	↓ Місяці –	↓ Роки	↓ Магазини – Міста	↓ Товари – Групи – Бренди				таблиці З вимірів відповідних рівнів

Матеріалізовані представлення

CREATE TABLE Fact_Sales

(part FK, supplier FK, customer FK, sales INT)

- Зберігають **наперед обчислені** значення сум адитивних метрик куба
- Агрегація виконується **лише на 1 рівні 1 ієрархії**, що експоненційно збільшує кількість при покритті всіх комбінацій рівнів вимірів
- Деякі СУБД підтримують автоматичне, інкрементне оновлення даних
- Запити користувачів мають бути коректно трансльовані

(part, supplier)

```
SELECT part, supplier, sum(sales)  
FROM Sales  
GROUP BY part, supplier
```

(part, customer)

```
SELECT part, customer, sum(sales)  
FROM Sales  
GROUP BY part, customer
```

(part)

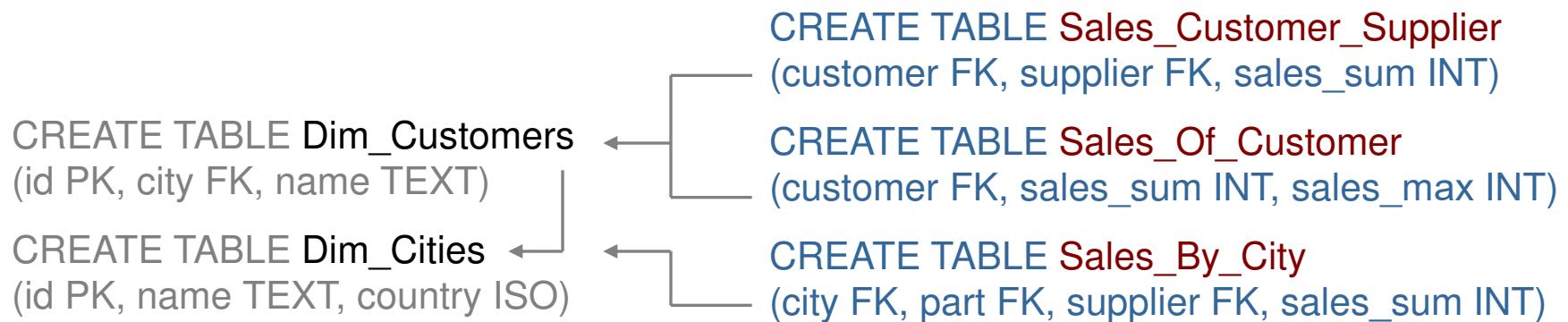
```
SELECT part, sum(sales)  
FROM Sales  
GROUP BY part
```

(supplier)

```
SELECT supplier, sum(sales)  
FROM Sales  
GROUP BY supplier
```

Таблиці агрегатів фактів

- Зберігають значення **суми та інших статистичних функцій** від метрик, отриманих на певному рівні деталізації
- Проектуються для пришвидшення **типових аналітичних задач** із кубом, з'ясовних при зборі потреб бізнес-аналізу або експлуатації
- Зв'язані зовнішніми ключами з вимірами відповідно до первинної таблиці фактів

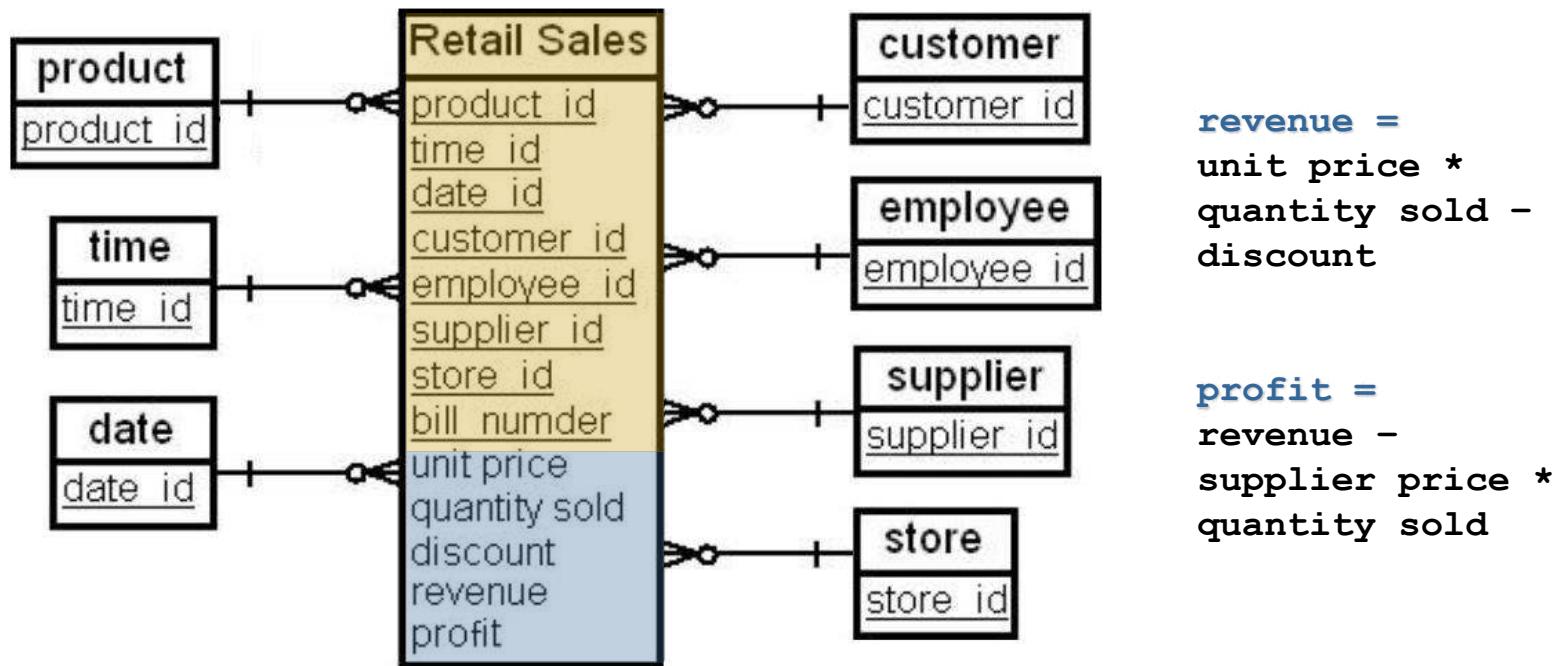


Індекси для таблиць фактів

- Комплексний первинний ключ на всіх колонках вимірів створює індекс (Unique), який працює **вкрай обмежно** при реальних операціях на кубі
- Додавання звичайних індексів (BTree) дозволяє пришвидшити фільтрацію, зріз та агрегацію **лише частково**, враховуючи комбінацію ключів
- Використання індексів бітової карти (Bitmap) суттєво зменшує розміри допоміжних структур та дає можливість ефективно комбінувати колонки вимірів при **фільтрації та агрегації**
- Створення індексів з'єднання (Join) продукує інверсний ключ до фактів, що дозволяє швидко **виконувати зрізи** по елементах вимірів
- Індекси вертикальної проекції (Column) на метриках пришвидшують **обчислення агрегатів**

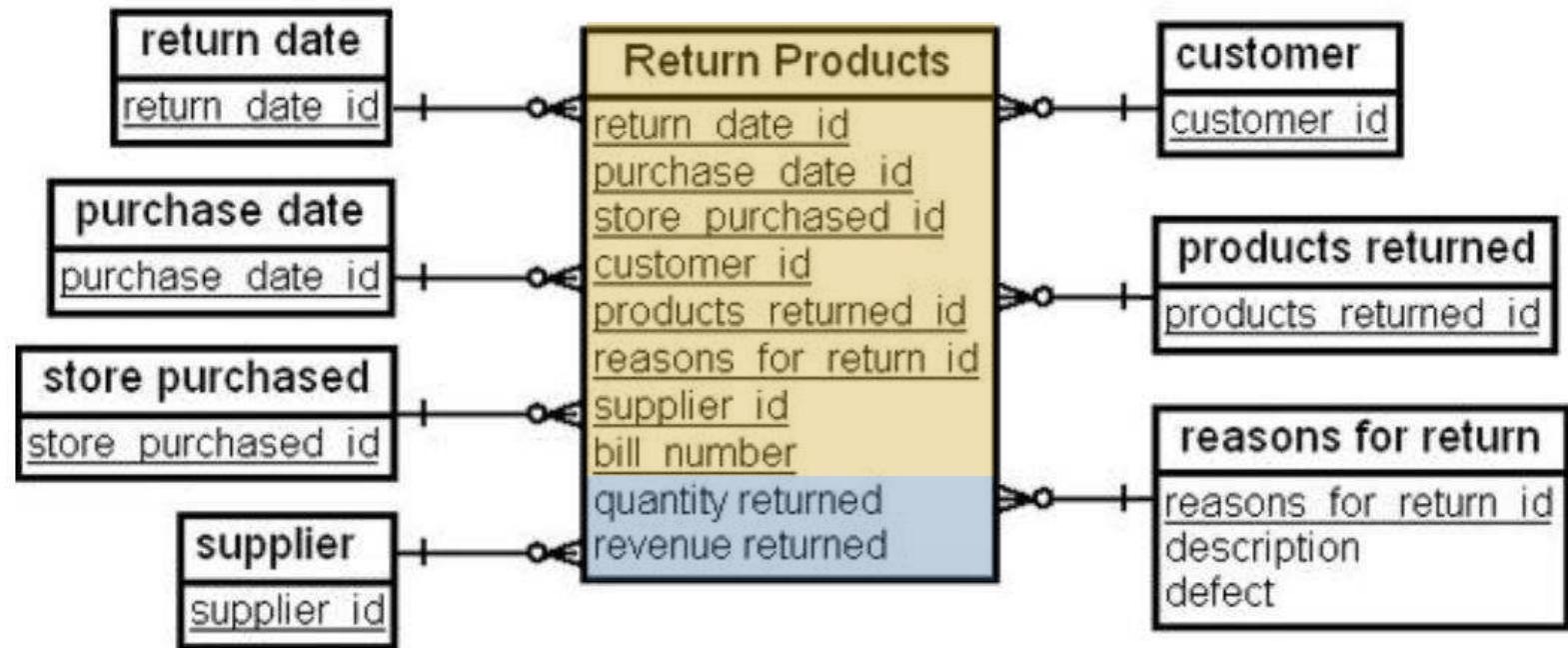
Роздрібна торгівля: приклад аналізу збуту товарів

- Факт описує продаж товару магазином згідно замовленню
- Метриками є характеристики позиції чеку і розраховані дані
- Вимірами є атрибути бізнес-події, необхідні для аналізу
- Дата та час розділені, номер чеку є виродженим виміром
- Постачальник доданий як окремий вимір, що вирішує проблему варіативності виміру товару в часі

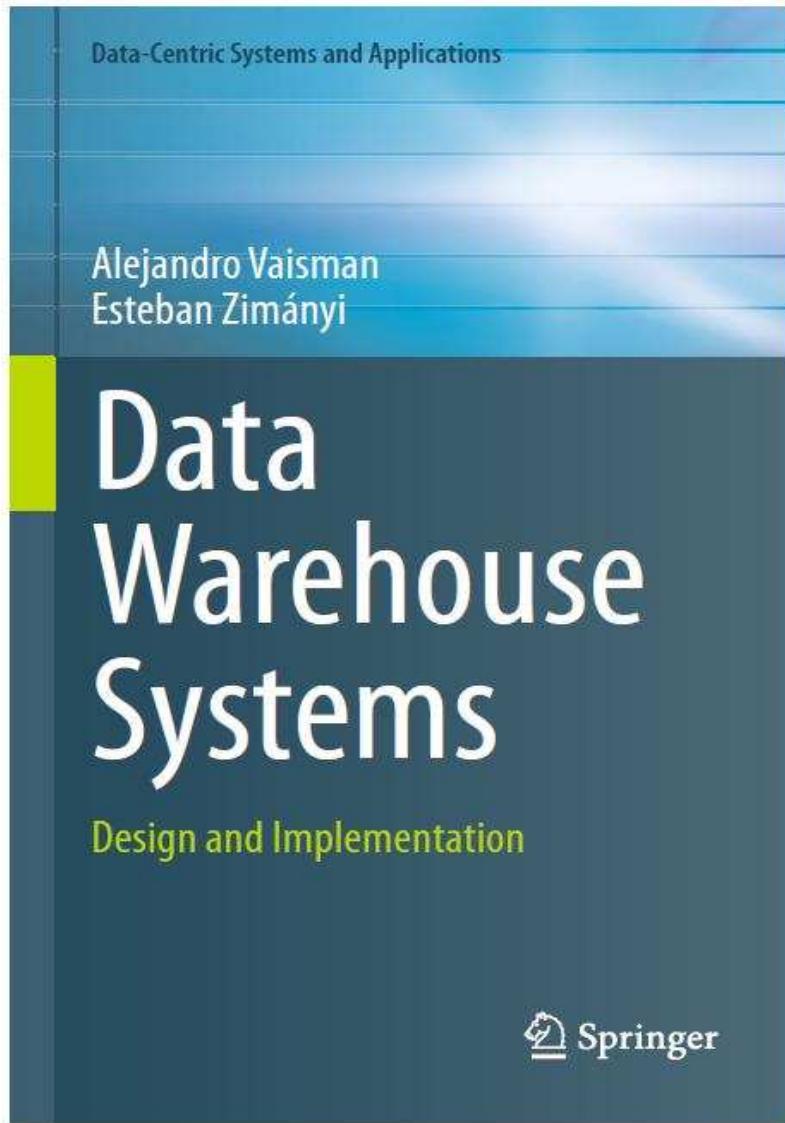


Роздрібна торгівля: приклад аналізу повернень товарів

- Факт описує повернення придбаного товару до магазину
- Метриками є важливі для аналізу повернень характеристики
- Вимірами є атрибути події, дати покупки та повернення
- Причина повернення зафіксована як вимір з ієрархією
- Поєднання за вимірами, що є спільними із фактом продажів, дає додаткові аналітичні показники бізнес-процесу



Дякую за увагу!



The Data Warehouse Toolkit

Third Edition

The Definitive Guide
to Dimensional
Modeling

Ralph Kimball
Margy Ross



СХОВИЩА ДАНИХ: Лекція №7

НУ “Львівська Політехніка”, кафедра ПЗ

Витягання, перетворення
та завантаження даних у сховище

Взаємодія із джерелами даних

■ Однорівнева архітектура

- Трансляція запитів до операційної бази
- Відсутнє дублювання даних
- Завжди актуальний стан
- Втручання в роботу OLTP-систем

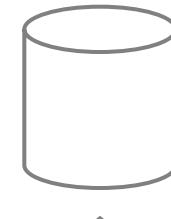
■ Накопичення сиріх даних

- Копіювання даних до “озера” у первинному вигляді
- Об’єднання схем і трансформація даних на стадії аналітичних запитів
- Складність систем бізнес-аналітики

■ Регулярне завантаження даних

- Внесення історичних даних із “озера”
- Отримання нових даних із джерел
- Окрема від OLTP та OLAP проміжна база
- Алгоритми синхронізації змін
- Узгодження та перевірка якості
- Додаткове складне обчислення

Операційні дані



Аналітичне
представлення



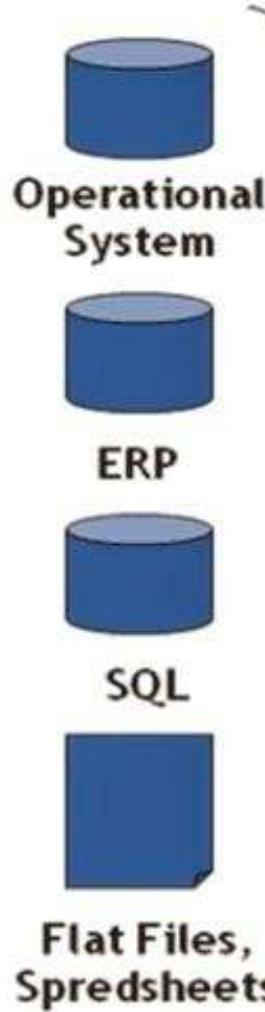
Звіти



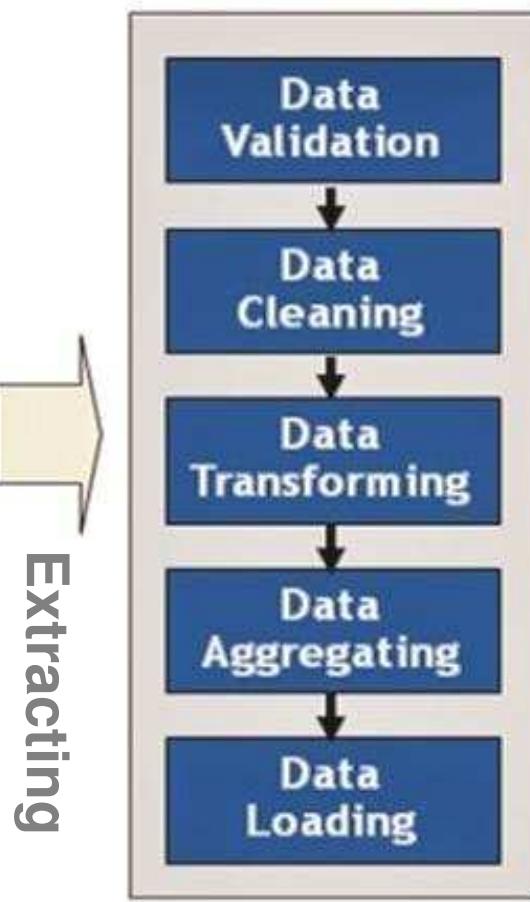
Куби

Внесення даних у сховище

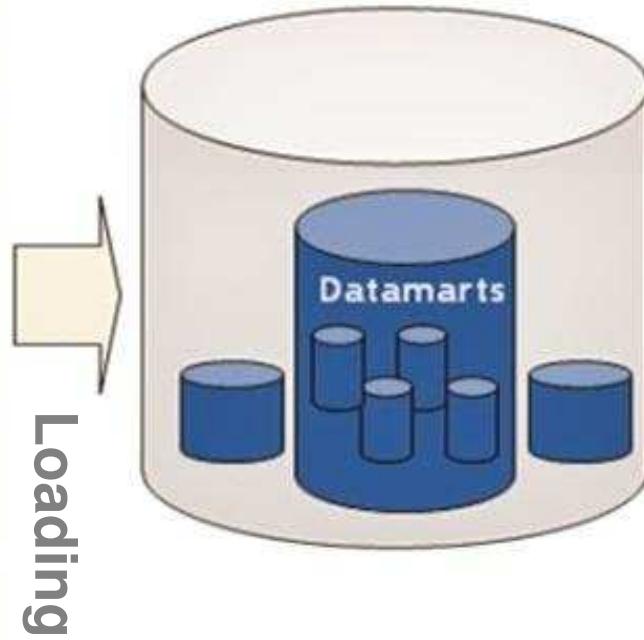
Data Sources



Staging Area



Data Warehouse



Основні стадії процесу ETL

- Відбір даних
 - Фільтрація
 - Консолідація
 - Валідація
 - Очищення
 - Дедублікація
- Трансформація
 - Денормалізація
 - Уніфікація значень
 - Обчислення метрик
 - Агрегація до рівня грануляції
- Завантаження
 - Встановлення змін
 - Актуалізація ключів
 - Оновлення даних вимірів
 - Внесення фактів
- Підтримка
 - Метадані сховища
 - Протоколювання
 - Обробка помилок
 - Планування повтору процесу

Складові частини: Getting Data into the Warehouse

- | | | |
|----|-------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------|
| 1. | Data Profiling — Explores a data source to determine its fit for inclusion as a source and the associated cleaning and conforming requirements | Logical Mapping
Data Profiling |
| 2. | Change Data Capture — Isolates the changes that occurred in the source system to reduce the processing burden | Isolate Changes
Landing, Staging |
| 3. | Extract System — Extracts and moves source data into the data warehouse environment for further processing | |

Result: Extracted tables with conversions

(of 34 total)

Складові частини: Cleaning and Conforming Data

4. **Data Cleaning System** — Implements data quality processes to catch quality violations
5. **Error Event Tracking** — Captures all error events that are vital inputs to data quality improvement
6. **Audit Dimension Creation** — Attaches metadata to each fact table as a dimension
7. **Data Deduplication** — Eliminates redundant members of core dimensions. May require integration across multiple sources and application of survivorship rules to identify the most appropriate version of a duplicate row
8. **Data Conformance** — Enforces common dimension attributes across conformed master dimensions and common metrics across related fact tables

Result: Cleaned tables and conformed dimensions

Складові частини: Preparing for Delivery and Presentation

- | | | |
|-----|----------------------------------------------------------------------------------------------------------------------------|---------------|
| 9. | Slowly Changing Dimension Manager | Time Variance |
| 10. | Surrogate Key Generator | |
| 11. | Hierarchy Dimension Manager | Bridge Tables |
| 12. | Special Dimensions Manager | Aggregates |
| 13. | Fact Table Builders | |
| 14. | Surrogate Key Pipeline | Hierarchies |
| 15. | Multi-Valued Bridge Table Builder | |
| 16. | Late Arriving Data Handler | |
| 17. | Dimension Manager — Centralized authority who prepares and publishes conformed dimensions data | |
| 18. | Fact Table Provider — Owns the administration of one or more fact tables and is responsible for maintenance and use | |
| 19. | Aggregate Builder — Builds and maintains aggregates to be used seamlessly with navigation technologies | |

Result: Ready fact and dimensions tables



Складові частини: Managing the Environment

- | | |
|-----------------------------------|----------------|
| 20. Multidimensional Cube Service | Run Constantly |
| 21. Data Propagation Manager | Reliability |
| 22. Job Scheduler | Availability |
| 23. Backup System | Manageability |
| 24. Recovery and Restart | Protection |
| 25. Version Control | Monitoring |
| 26. Version Migration | |
| 27. Workflow Monitor | |
| 28. Sorting System | |
| 29. Lineage and Dependency | |
| 30. Problem Escalation | |
| 31. Parallelizing and Pipelining | |
| 32. Security System | |
| 33. Compliance Manager | |
| 34. Metadata Repository | |

Допоміжні бази даних

■ Landing – raw data format

- Snapshots of the original data from the sources (performing complete extraction, incremental, change tracking or dates range)
- Does not persist the data between landing process runs

■ Staging – pre-warehouse data

- Transforms and aggregates data of the landing store joined with the historical details of the business entities
- Prepares data to be merged into the warehouse database
- Reduces load on the source systems during the complex transformation routines
- Allows rapid failure recovery, because the data does not need to be extracted a second time

■ Persistent – life-cycle temporal data

- Stores all historical changes of the processed source entities
- Allows to reload the data warehouse with full history (due to a change in logic, model or mistakes)
- Provides trace and audit abilities of the data loading process

Правила обробки даних

- Створення діаграм потоків даних від джерел, визначення релевантних атрибутів, правил перетворень та агрегації, операцій оновлення для всіх таблиць сховища
- Розробка алгоритмів очищення та критеріїв валідації даних, що надходять від джерел
- Створення документу, який визначає відповідність моделей даних з урахуванням зв'язків між сущностями



- Розробка процедур заповнення та верифікації таблиць-вимірів, первинного внесення та періодичного оновлення даних таблиць-фактів, агрегатів
- Особливості періодичної синхронізації відповідно до структури даних та потреб бізнес-аналізу
- Визначення вимог з питань обмеження доступу

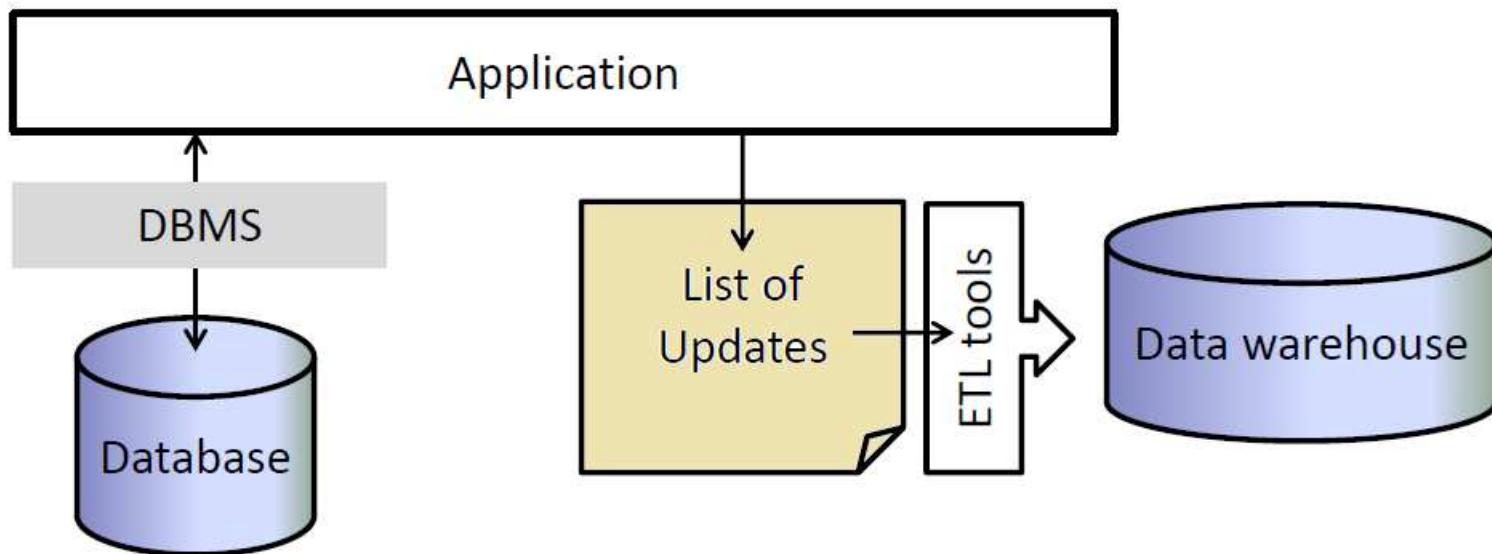
Опрацювання джерел даних

- Обсяги синхронізації даних
 - Повна, за періодом часу
 - Інкрементальна, реального часу
- Можливості інформаційних систем
 - Підтримка довільних запитів
 - Отримання знімків, різниці даних
- Метод виявлення оновлень
 - Функція на рівні додатків
 - Тригер при зміні даних
 - Обробка журналів транзакцій
 - Збереження міток часу
 - Повне порівняння даних

Application-Assisted Synchronization

Update of data warehouse deeply integrated into the application software

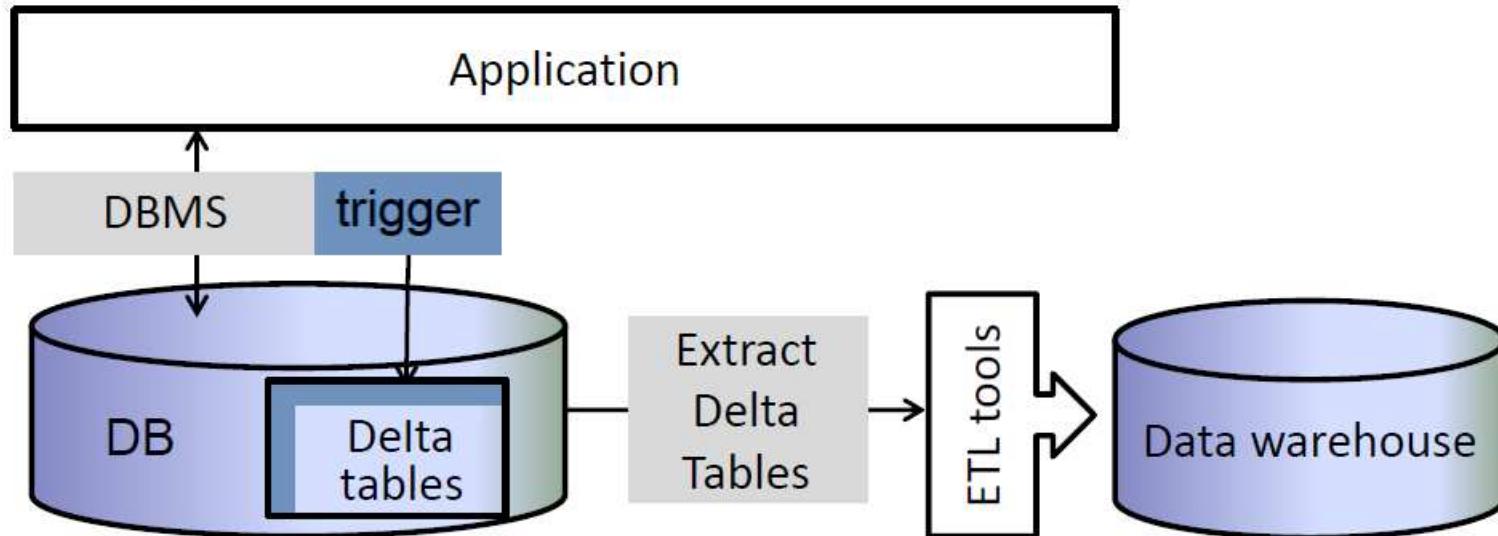
- Immediate
- Requires adaptation of existing applications
- Hard to maintain



Trigger-Based Synchronization

Closely related to application-based

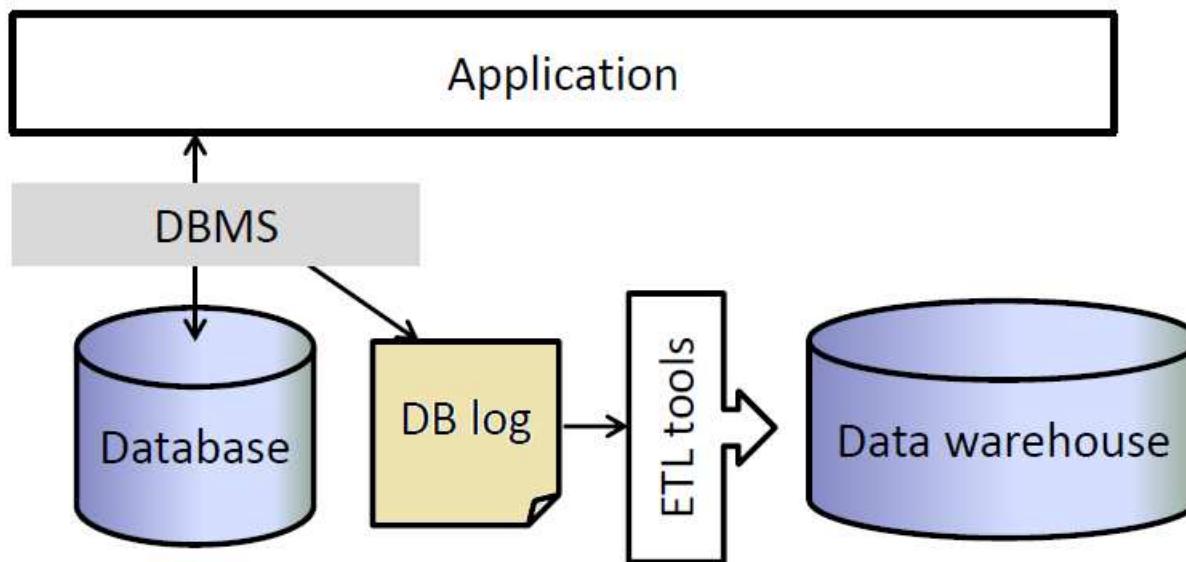
- Triggers to store updates
- Huge performance hit for database
- More transparent for application layer
- Only possible for data maintained in database



Log-Based Synchronization

Many database systems keep change logs

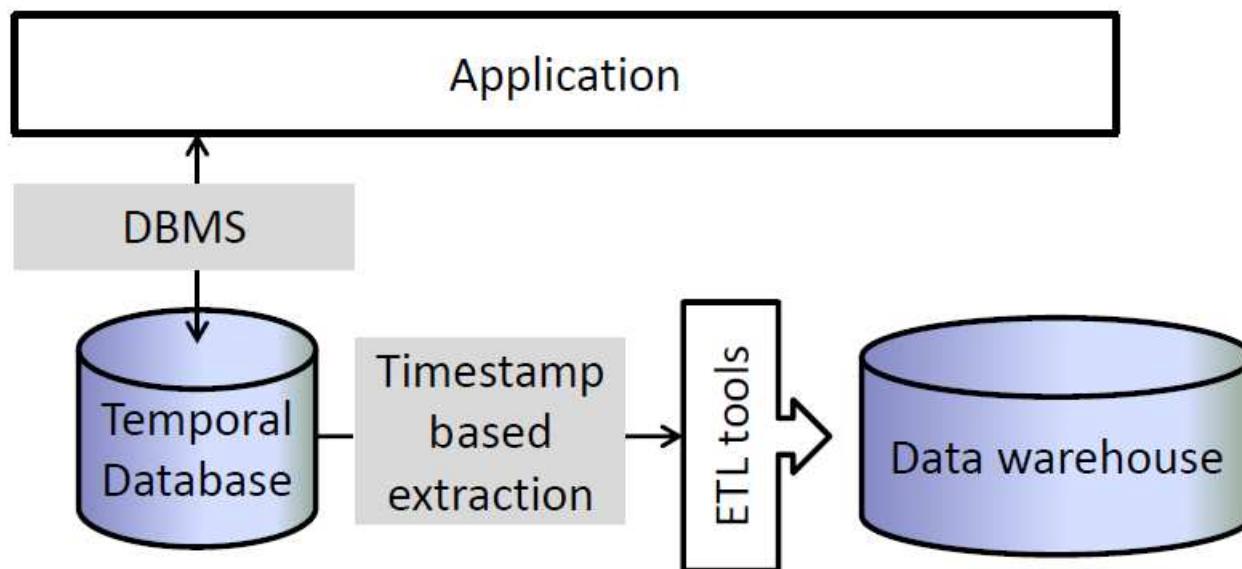
- To ensure durability; in-between checkpoints all transactions to the database are logged
- Use these logs for updating data warehouse
- Transparent to the user



Timestamp-Based Synchronization

Operational database can be changed to keep whole history

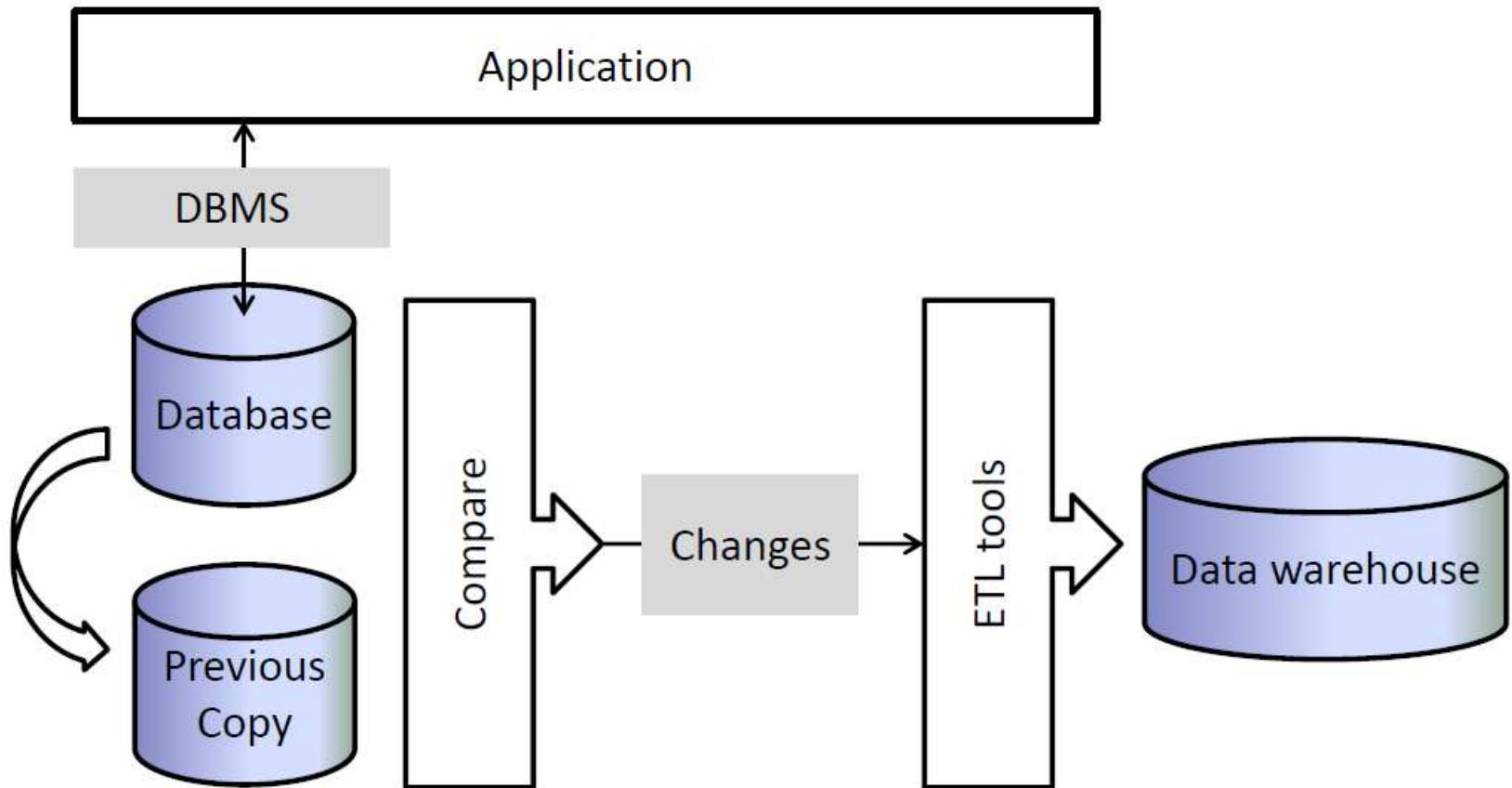
- One timestamp per tuple
- Fine-grained full temporal database (start-end)



Difference-Based Synchronization

Easy, works even for flat files

- Does not capture intermediate changes



Дані у сховищі повинні бути:

- точними – дані повинні містити правильні кількісні значення метрик або давати пояснення, чому неможливо мати такі значення
- повними – користувачі системи БА повинні знати, що мають доступ до всіх релевантних даних
- узгодженими – не допускаються ніякі суперечності в даних: агрегати повинні точно відповідати детальним даним
- унікальними – одні й ті ж об'єкти предметної області повинні мати однакові назви та ідентифікуватися у сховищі даних одинаковими ключами
- актуальними – користувачі системи бізнес-аналітики повинні знати, з якою частотою оновлюються дані (тобто на яку дату дані є дійсними)

Процедурами очищення можуть бути:

- конвертація і нормалізація даних
(приведення тексту до однакового кодування, однакові формати дати і т. д.)
- стандартизація написання назв, подання адрес, усунення дублікатів
- стандартизація найменувань таблиць, індексів і т.д.
- очищення, яке базується на бізнес-правилах предметної області
- створення міток статусу фактів в таблицях-вимірах (нормальний, ненормальний, неможливий, такий, що виходить за межі, аналізований чи ні і т.д.)
- уніфікація використання NULL-значень
- маркування фактів зі зміненим статусом (наприклад, покупець повернув товар)

Unstandardized Units	Outliers and Errors
Missing Values	Incompatible Data
Misspelled Entries	Duplicate Records
Time-Changing Values	Out of Ranges

Відкидання некоректних даних, уніфікація назв, приведення типів:

Місто	Рік	Кілкість
Львів	1991	100,00
Київ	1991	200
Харків, Україна	123456	300
Харьков, Украина	93	400
Днепропетровск	1993	500



Місто	Рік	Кілкість
Львів	1991	100
Київ	1991	200
Харків	1993	400
Дніпро	1993	500

Спрощення назв, врахування часу, виявлення невідомих значень:

Прізвище	Ім'я	Стать	Вік
Шевченко	Тарас	чол.	24
Петренко Микола			23
Тарасенко	Ірина	жін.	22
б/п	Андрій	н/д	21
Богдан		чол.	20
Олена Іванівна		жін.	0



Прізвище, ім'я	Стать	Рік нар.
Шевченко Тарас	чол.	2000
Петренко Микола	NULL	2001
Тарасенко Ірина	жін.	2002
Андрій	NULL	2003
Богдан	чол.	2004
Олена Іванівна	жін.	NULL

Врахування різних назв тих самих об'єктів, збереження посилань:

Підрозділ	Код №1	Код №2
"Склад"	00-100	СК1
Магазин	00-200	
Офіс		ОФ1
Office	00-300	



Назва підрозділу	Код	Джерело
Не встановлено	NULL	NULL
Склад	00-100	№1
Склад	СК1	№2
Магазин	00-200	№1
Офіс	ОФ1	№2
Офіс	00-300	№1

Перетворення у формат сховища

- ❑ Конвертація типів даних
 - символльні рядки (регистр, з'єднання, єдиний шаблон)
 - формати дати, часу, координат
 - числа із комами та пробілами
 - дискретизація значень атрибутів
- ❑ Комбінування різних джерел
 - об'єднання за назвами при відсутності єдиного коду
 - заміщення кодів значеннями з каталогу назв
- ❑ Обчислення значень метрик
 - різниця між мітками часу
 - відсотки, коефіцієнти, відстані
 - розрахунки за формулами
- ❑ Агрегати для фактів знімків
 - накопичена сума, середнє, кількість за період
 - статистичні функції на наборах

Конвертація числового формату, встановлення зовнішніх ключів:

№ транзакції	№ терміналу	Дата	Сума
1	T1	15.03.2024	5,000.10
1	T2	15.03.2024	500.20
2	T2	15.03.2024	1,000.00
1	T3	18.03.2024	2,000.00
2	T1	18.03.2024	3,000.00



Код	Дата	Сума
T1/1	Д24_03_15	5000.1
T2/1	Д24_03_15	500.2
T2/2	Д24_03_15	1000
T3/1	Д24_03_18	2000
T1/2	Д24_03_18	3000

Обчислення значення часу, визначення типу за текстовим описом:

Початок	Кінець	Дата	Коментар до операції
10:00	11:00	15.03.2024	Підключення доступу до мережі
14:00	14:30	15.03.2024	Налаштування комп'ютерів
12:00	14:00	18.03.2024	Встановлення програм
15:00	15:30	18.03.2024	Встановлення з'єднання із мережею
17:00	20:00	01.04.2024	Прибирання приміщенъ



Тривалість	Дата	Тип операції
60	Д24_03_15	Мережа
30	Д24_03_15	Налаштування
120	Д24_03_18	Налаштування
30	Д24_03_18	Мережа
180	Д24_04_01	Клінінг

Використання існуючих первинних ключів сховища:

Код	Дата	День
Д24_01_10	10 січня 2024	срд
Д24_03_15	15 березня 2024	птн
Д24_03_18	18 березня 2024	пнд
Д24_04_01	1 квітня 2024	пнд

Завантаження даних вимірів

- Актуалізація таблиць поточними даними
 - Створення нових значень по діапазонах
 - Генерація значень статичних вимірів
 - Внесення рядків/ключів відповідно до схеми
- Перевірка збалансованості та множинності відношень в ієрархії
 - Актуалізація таблиць-мостів
 - Помилка у випадку виявлення змін/зв'язків, що не реалізовані у схемі сховища
- Виявлення змін атрибутів із моменту останньої синхронізації
 - Використання міток часу або порівняння
 - Обробка відповідно до типу виміру

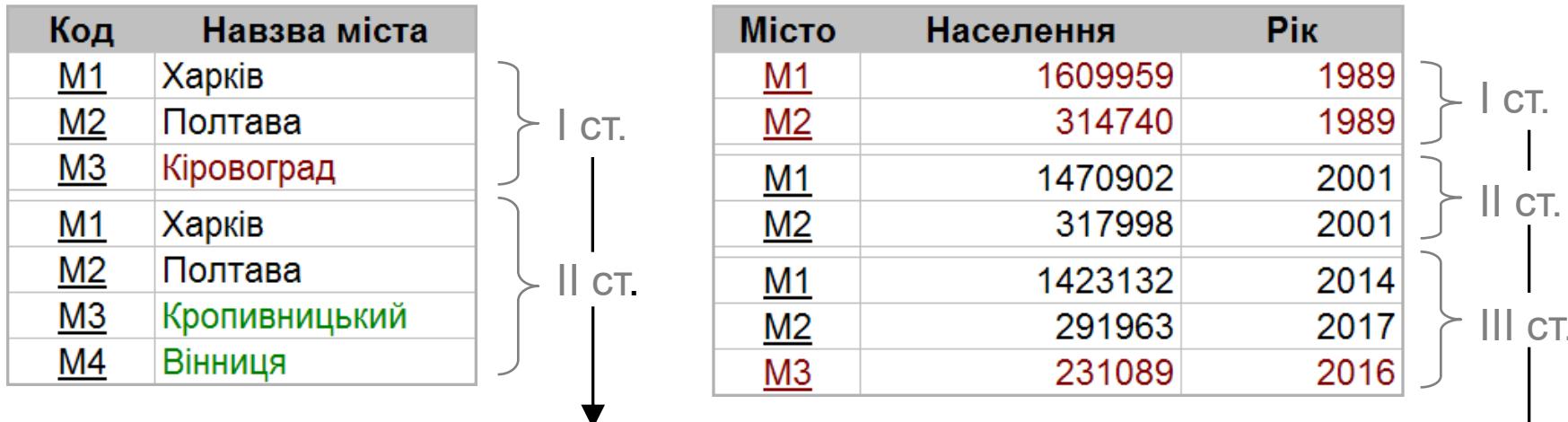
Оновлення таблиць повільно змінних вимірів

	Type 1 SCD	Type 2 SCD	Type 3 SCD
Initial Load	Initial Load	Initial Load, Tracking Established	Initial Load, History Columns Established
Incremental Load	<p>New Data</p> <pre>graph TD; ND[New Data] --> MD{Match}; MD -- No --> LNR[Load New Records]; MD -- Yes --> ROR[Replace Old Records];</pre> <p>Change detection is optional</p>	<p>New Data</p> <pre>graph TD; ND[New Data] --> MD{Match}; MD -- No --> LNR[Load New Records]; LNR --> MOV[Mark Old Version Obsolete]; MD -- Yes --> CD{Change Detection}; CD --> ANV[Add New Version]</pre>	<p>New Data</p> <pre>graph TD; ND[New Data] --> MD{Match}; MD -- No --> LNR[Load New Records]; LNR --> CD{Change Detection}; MD -- Yes --> CD{Change Detection}; CD --> MCR[Modify Changed Records]</pre>

Завантаження даних фактів

- Внесення рядків, збережених у проміжній базі відповідно до методу синхронізації
- Оновлення фактів, які були змінені або видалені у первинній базі
- Оновлення фактів-агрегатів, що покривають період від останньої синхронізації (повний перерахунок)
- Якщо є факти без значень метрик, то необхідно дослідити коректність процесу
- Якщо змінюються всі рядки таблиці, то слід виправити схему сховища

Зміна рядків таблиць первинного джерела даних із часом (стадії):



Оновлення таблиці виміру при додаванні міст та зміні назви:

Місто	Поточна назва	Внесено	Попередня назва	Оновлено
M1	Харків	24.08.1991		
M2	Полтава	24.08.1991		
M3	Кропивницький	24.08.1991	Кіровоград	16.07.2016
M4	Вінниця	16.07.2016		

Внесення рядків до таблиці фактів, що зберігає динаміку населення:

Внесено	Місто	Населення	Зміна	Років
31.10.2002	M1	1 470 902	-8,64%	12
31.10.2002	M2	317 998	+1,04%	12
24.08.2018	M1	1 423 132	-3,25%	13
24.08.2018	M2	291 963	-8,19%	16

Особливості завантаження до реляційної бази даних

- Спочатку внесення даних вимірів, потім – фактів
- Первинне завантаження – циклами по роках
- Інкрементальне завантаження поділяється за частотою на оперативне, проміжне, щодобове
- Часові мітки на початку та для відновлення
- Актуалізація матеріалізованих представлень
- Типові шляхи пришвидшення:
 - Відключення перевірки цілісності посилань за зовнішніми ключами для таблиць фактів
 - Видалення індексів перед масивним оновленням
 - Пониження рівня ізоляції транзакцій СУБД
 - Паралельна обробка різних потоків даних
 - Безперервне внесення транзакційних фактів
 - Пакетна обробка рядків з даними

Основні функції програмного забезпечення для інтеграції даних

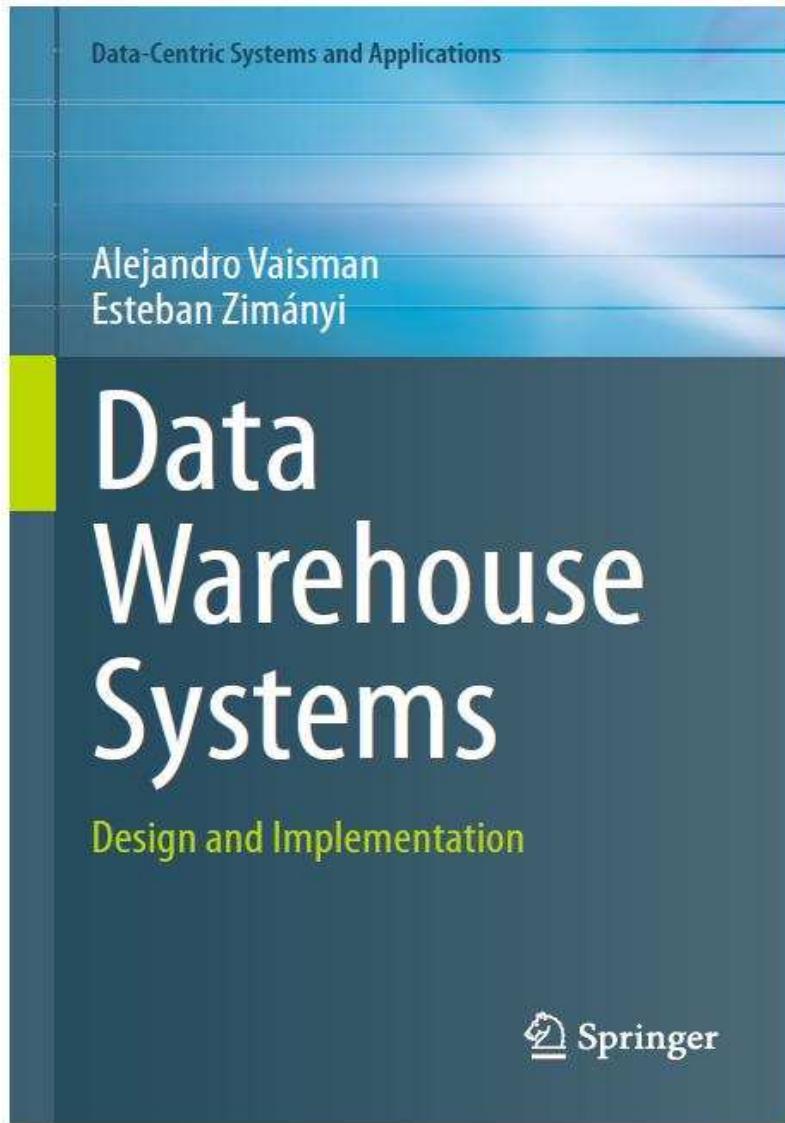
- Доступ до джерел та схем даних
- Підтримка власної робочої бази
- Трансформація у цільовий формат
- Забезпечення рівня якості даних
- Повне первинне завантаження
- Відстеження та внесення змін
- Протоколювання операцій та помилок
- Активація за розкладом або по запиту
- Відновлення після переривання
- Готовність до розширення моделей
- Прозорість процедур трансформації



Extracting, Transformation, Loading

- 19 ETL Subsystems and Techniques
 - Round Up the Requirements
 - Business Needs
 - Compliance
 - Data Quality
 - Security
 - Data Integration
 - Data Latency
 - Archiving and Lineage
 - BI Delivery Interfaces
 - Available Skills
 - Legacy Licenses
 - The 34 Subsystems of ETL
 - Extracting: Getting Data into the Data Warehouse
 - Subsystem 1: Data Profiling
 - Subsystem 2: Change Data Capture System
 - Subsystem 3: Extract System
 - + Cleaning and Conforming Data
 - + Delivering: Prepare for Presentation
 - + Managing the ETL Environment
 - Summary
- 20 ETL System Design and Development Process and Tasks
 - ETL Process Overview
 - Develop the ETL Plan
 - Step 1: Draw the High-Level Plan
 - Step 2: Choose an ETL Tool
 - Step 3: Develop Default Strategies
 - Step 4: Drill Down by Target Table
 - Develop the ETL Specification Document
 - Develop One-Time Historic Load Processing
 - Step 5: Populate Dimension Tables with Historic Data
 - Step 6: Perform the Fact Table Historic Load
 - Develop Incremental ETL Processing
 - Step 7: Dimension Table Incremental Processing
 - Step 8: Fact Table Incremental Processing
 - Step 9: Aggregate Table and OLAP Loads
 - Step 10: ETL System Operation and Automation
 - Real-Time Implications
 - Real-Time Triage
 - Real-Time Architecture Trade-Offs
 - Real-Time Partitions in the Presentation Server
 - Summary

Дякую за увагу!



The Data Warehouse Toolkit

Third Edition

The Definitive Guide
to Dimensional
Modeling

Ralph Kimball
Margy Ross



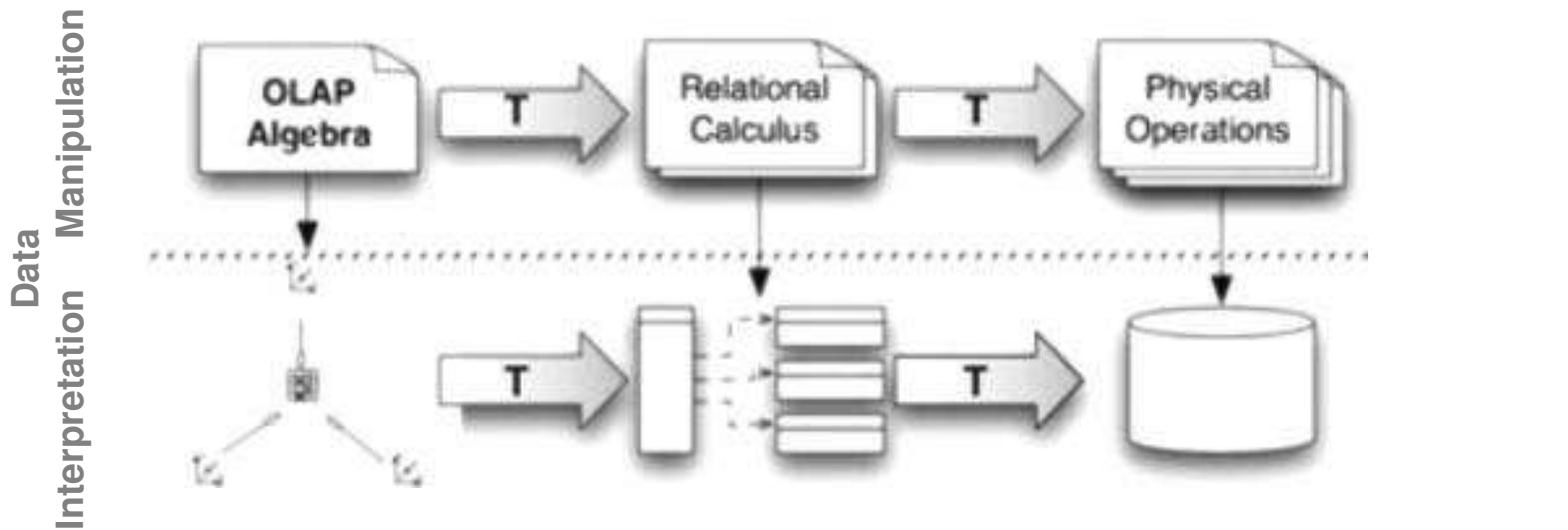
СХОВИЩА ДАНИХ: Лекція №8

НУ “Львівська Політехніка”, кафедра ПЗ

Вибірка даних із реляційної
моделі сховища

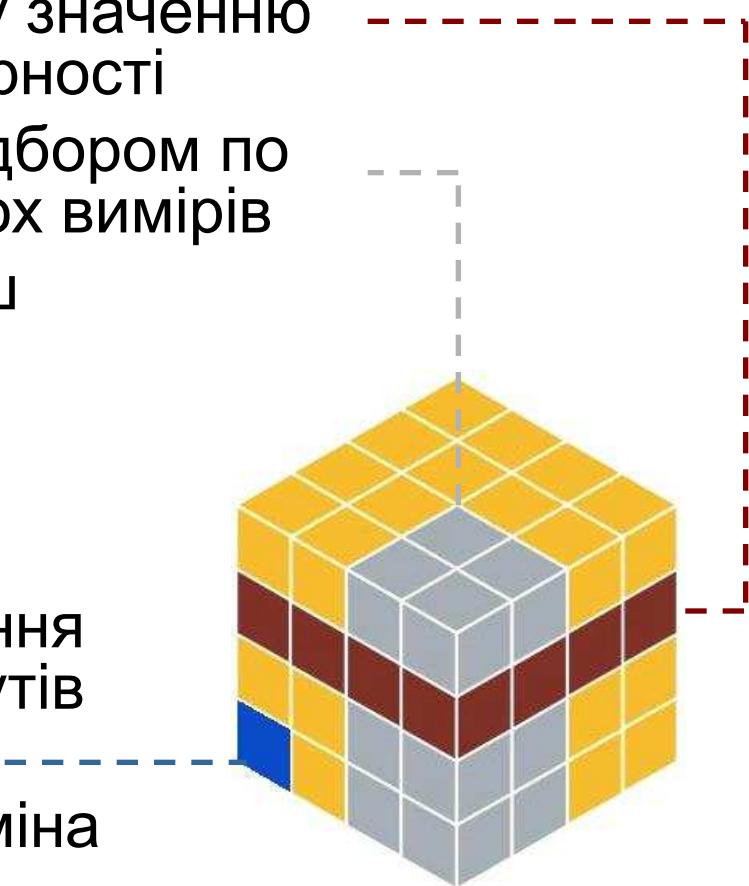
Трансляція запитів на рівень обробки даних

- Користувачі систем бізнес-аналітики (BI Tools) оперують даними на концептуальному рівні
 - створення дашбордів, графіків, агрегованих звітів
 - об'єднання, обчислення показників фактів
 - виконання навігації по багатовимірній моделі
- Клієнтські програми (Front End) формують запити до OLAP-серверу згідно описаних моделей
- Процесор запитів (Query Engine) транслює запити до СУБД у відповідності до метаданих сховища
 - фільтрація, агрегація, групування даних
 - з'єднання таблиць фактів та вимірів



Операції презентації на кубі

- **SLICE** – зріз куба по одному значенню виміру із пониженням розмірності
- **DICE** – формування кубу відбором по діапазонах значень декількох вимірів
- **ROLL UP** – перехід на більш загальний рівень виміру із агрегацією значень метрик
- **DRILL DOWN** – перехід на детальний рівень виміру
- **DRILL THROUGH** – отримання значень метрик і всіх атрибутів вимірів найнижчого рівня
- **PIVOT** – поворот по осях, зміна форми виведення даних



Оператори багатовимірної алгебри

- Трансформація кубу, згортка або розгортка:
 - **PROJECTION** – обрання множин метрик та вимірів на відповідних рівнях
 - **LEVEL CHANGE** – зміна рівня ієрархії за виміром
 - **BASE CHANGE** – додавання або видалення виміру
- Метрики та множина вимірів без змін:
 - **SELECTION** – фільтрація по значеннях вимірів
- Обробка на перетині вимірів кубів:
 - **COMBINE** – обчислення нових значень метрик
 - **JOIN** – об'єднання різних метрик

A(5[1] x 4[1])	Чернівці	Тернопіль	Львів	Краків	Люблін
Микола	1	2	3	2	4
Олександр	2	3	4	3	3
Анна	3	3	3	3	2
Марія	4	2	4	2	1

PROJECTION

O(n)	-
-	54

↓ *SELECTION (дія DICE)*

LEVEL CHANGE (дія ROLL UP)

B(3[1] x 4[1])	Львів	Краків	Люблін
Микола	3	2	4
Олександр	4	3	3
Анна	3	3	2
Марія	4	2	1

C(2[2] x 4[1])	Україна	Польща
Микола	3	6
Олександр	4	6
Анна	3	5
Марія	4	3

↓ *BASE CHANGE (дія ROLL UP)*

JOIN (дія DRILL ACROSS)

D(4[1], m)	-
Микола	9
Олександр	10
Анна	8
Марія	7



E(4[1], n)	-
Микола	30
Тарас	40
Анна	50
Олена	60

F(6[1], m, n)	-
Микола	9; 30
Олександр	10; 0
Тарас	0; 40
Анна	8; 50
Марія	7; 0
Олена	0; 60

Операції комбінування кубів

- **DRILL ACROSS** – формування кубу по всім спільними вимірам із об'єднанням різних метрик
- **UNION** – додавання значень ідентичних метрик кубів, заданих на спільних вимірах
- **INTERSECTION** – обчислення різниці між більшим та меншим значеннями метрик
- **DIFFERENCE** – збереження метрики, яка має більше числове значення

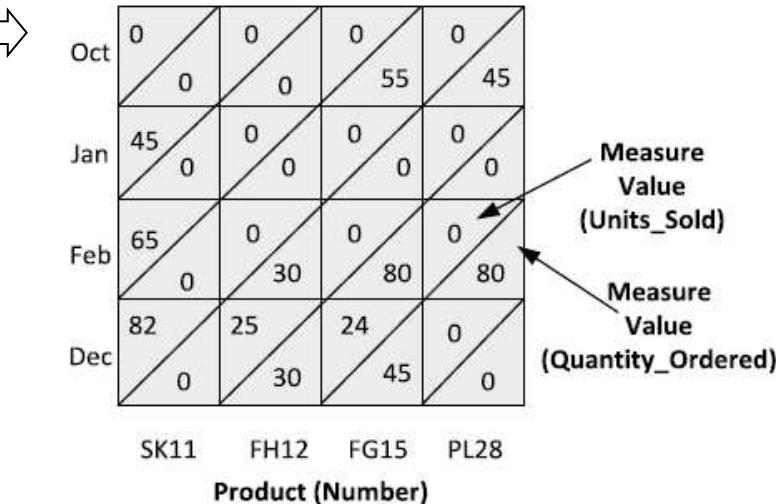
$$M = \{ Pa, Qb \}$$

$$M = Ma + Mb$$

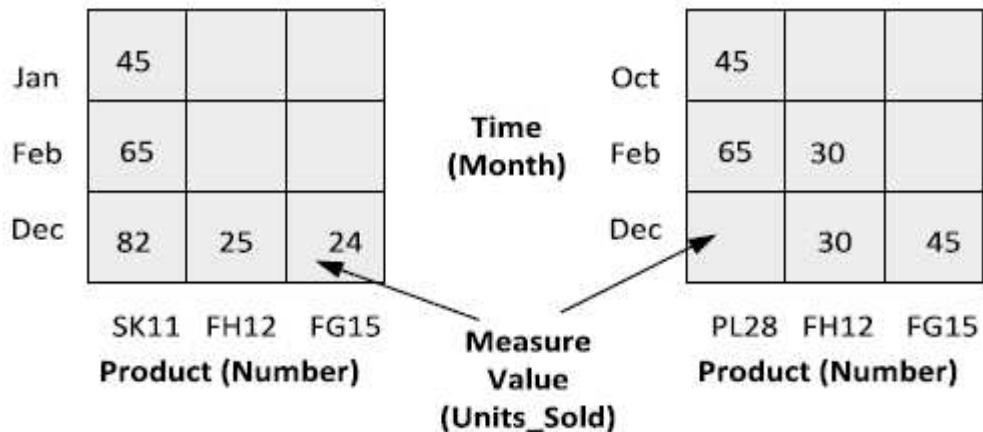
$$M = | Ma - Mb |$$

$$M = \max(Ma, Mb)$$

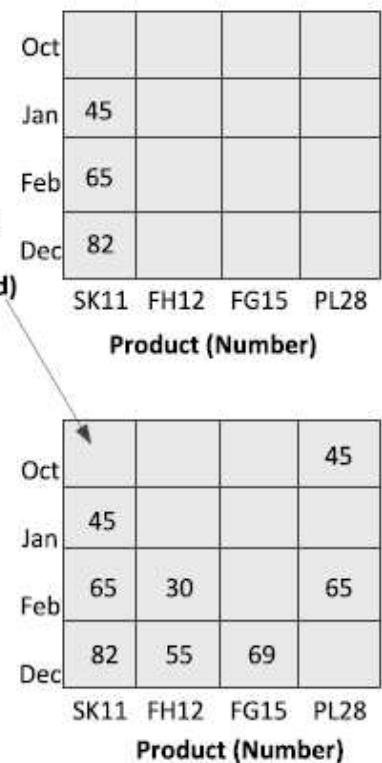
Операція DRILL ACROSS



Операція DIFFERENCE



Операція UNION



Перехід від кубів до таблиць

- Багатовимірний куб презентує дані всіх рівнів
 - операція проекції трактується як вибірка
 - у заголовках таблиці розміщені значення вимірів
- Таблиця фактів має дані одного рівня ієархії
 - проекція виконується як групування та обчислення суми по метрикам
 - селекція збирає ключі вимірів та застосовує їх як умову фільтрації рядків
 - окремі таблиці обираються для обчислених агрегатів

	c1	c2	c3	Total
p1	100	105	100	305
p2	70	60	40	170
p3	30	40	50	120
Total	200	205	190	595

SELECT
SUM(...)
GROUP BY

ProductKey	CustomerKey	SalesAmount
p1	c1	100
p1	c2	105
p1	c3	100
p2	c1	70
p2	c2	60
p2	c3	40
p3	c1	30
p3	c2	40
p3	c3	50

Вибірка із таблиці фактів

- Агрегатні функції не потрібні, якщо в проекцію включені всі нижні рівні всіх вимірів

```
SELECT Sales.*
```

```
FROM Sales
```

```
WHERE CustomerKey IN ('c1', 'c2', 'c3')
```

+ селекція за
виміром

- Функція суми обирається для метрики, якщо вона адитивна за вимірами проекції

```
SELECT CustomerKey, SUM(SalesAmount)  
FROM Sales  
GROUP BY CustomerKey
```

+ мін. або макс.
значення

- Підрахунок кількості рядків застосовується виключно для транзакційних фактів

```
SELECT CustomerKey, ProductKey, COUNT(*)  
FROM Sales  
GROUP BY CustomerKey, ProductKey
```

+ статистичні
функції від кількості

Використання комбінування груп

ProductKey	CustomerKey	SalesAmount
p1	c1	100
p2	c1	70
p3	c1	30
NULL	c1	200
p1	c2	105
p2	c2	60
p3	c2	40
NULL	c2	205
p1	c3	100
p2	c3	40
p3	c3	50
NULL	c3	190
NULL	NULL	595
p1	NULL	305
p2	NULL	170
p3	NULL	120

Отримання додаткових проекцій з кубу

```
SELECT ProductKey, NULL, SUM(SalesAmount)
FROM Sales
GROUP BY ProductKey
UNION
SELECT NULL, CustomerKey, SUM(SalesAmount)
FROM Sales
GROUP BY CustomerKey
```

Специфікація наборів груп GROUPING SETS

```
SELECT ProductKey, CustomerKey, SUM(SalesAmount)
FROM Sales
GROUP BY GROUPING SETS((ProductKey, CustomerKey),
(ProductKey), (CustomerKey), ())
```

Синтаксичні скорочення ROLLUP та CUBE

```
SELECT ProductKey, CustomerKey, SUM(SalesAmount)
FROM Sales
GROUP BY CUBE(ProductKey, CustomerKey)
```

Використання віконних функцій

```
SELECT ProductKey, CustomerKey, SalesAmount, MAX(SalesAmount) OVER  
(PARTITION BY ProductKey) AS MaxAmount
```

Product Key	Customer Key	Sales Amount	Max Amount
p1	c1	100	105
p1	c2	105	105
p1	c3	100	105

Клауза PARTITION BY для включення даних з вищого рівня
(максимальне по продукту)

```
SELECT ProductKey, CustomerKey, SalesAmount, ROW_NUMBER() OVER  
(PARTITION BY CustomerKey ORDER BY SalesAmount DESC) AS RowNo
```

Product Key	Customer Key	Sales Amount	RowNo
p1	c1	100	1
p2	c1	70	2
p3	c1	30	3

Клауза ORDER BY для впорядкування результатів
(місце за об'ємом продажів)

```
SELECT ProductKey, Year, Month, SalesAmount, AVG(SalesAmount) OVER  
(PARTITION BY ProductKey ORDER BY Year, Month ROWS 2 PRECEDING) AS MovAvg
```

Product Key	Year	Month	Sales Amount	MovAvg
p1	2011	10	100	100
p1	2011	11	105	102.5
p1	2011	12	100	101.67

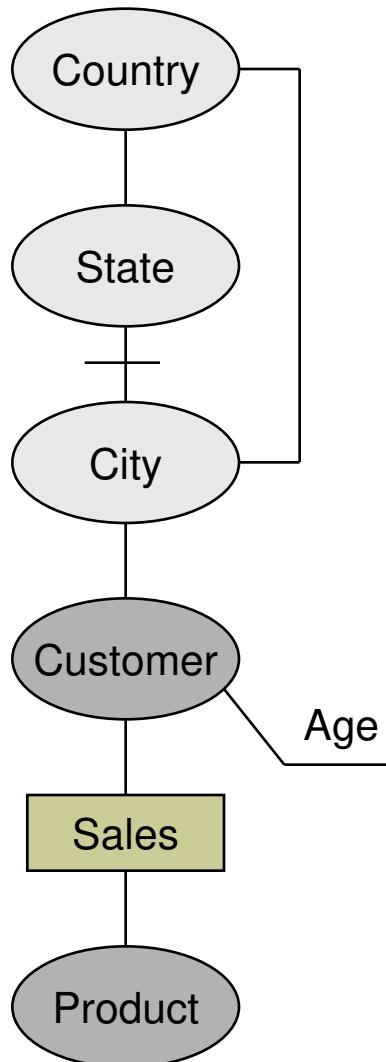
Клауза ROWS для обмеження вікна сканування
(середнє за попередні місяці)

Призначення даних вимірів

- Таблиці накопичують характеристики процесів, виявлені під час функціонування сховища, а не опис самих процесів (об'єктів, подій)
 - Простір назв, що відповідають множинам первинних ключів таблиці, найближчої до факту
 - Додаткові атрибути для пошуку
 - Поточні дані можуть не мати зв'язків з фактами
- Нормалізовані таблиці “сніжинки” з'єднуються за реляційними ключами відповідно до схеми
- Агрегування не передбачено для таблиць вимірів
 - Запит на отримання аналітичних даних має завжди включати з'єднання до таблиць фактів
 - Використання атрибутів вимірів є помилковим при виведенні будь-яких обчислювальних метрик

Date Key	Day	Month	Year
d1	31.03.25	3.2025	2025
d2	01.04.25	4.2025	2025
d3	02.04.25	4.2025	2025

Вибірка із таблиці вимірів



Нижній рівень має унікальні первинні ключі для кожного елементу виміру

```
SELECT Customer.Key, Customer.Name  
FROM Customer  
WHERE Customer.Age > 25  
ORDER BY Customer.Name
```

+ фільтрація за атрибутом

Використовується DISTINCT, щоб отримати значення проміжних рівнів

```
SELECT DISTINCT Customer.City  
FROM Customer  
ORDER BY Customer.City
```

+ запит на вибірку ключів

Заміна або балансування пропущеного значення проміжного рівня

```
SELECT COALESCE(Customer.State, 'Other') ...  
SELECT COALESCE(Customer.State, Customer.Country) ...
```

Пропущені значення виміру

При проекції враховуються пропущені значення (необов'язкові гілки ієрархії), щоб загальна сума була однаковою на всіх рівнях презентації даних

Country	State	City	QTY
USA	Texas	Dallas	1
Belgium	-	Brussels	5
Belgium	-	Antwerp	5
Germany	-	Berlin	3
Germany	-	Dresden	8

$$\sum = 22$$

підстановка



балансування

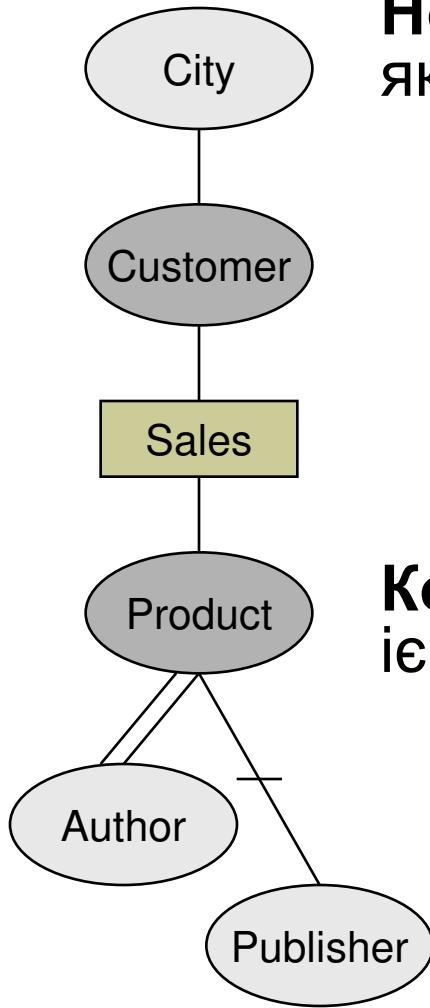
State	QTY
Texas	1
Other	21

State	QTY
Texas	1
Belgium	10
Germany	11

Запит до таблиць фактів і вимірів

SELECT SUM(fact.measure), dim1.bottom_level, dim2.top_level	Агрегатна функція від метрики та стовпці із назвами елементів вимірів
FROM fact INNER JOIN dim1 ON ... INNER JOIN dim2 ON ...	Зовнішній ключ задано для обов'язкових вимірів при з'єднанні
FROM fact LEFT JOIN dim_optional ON ...	Вимір необов'язковий, має бути заміна назви для пустого значення
FROM dim1 CROSS JOIN dim2 LEFT JOIN fact ON ...	Декартів добуток для презентації всіх значень з таблиць вимірів
WHERE dim1.key IN (...)	Зріз даних на визначених рівнях за ключами або атрибутами
GROUP BY dim1.bottom_level, dim2.top_level	Групування по значеннях вимірів відповідно до операції проекції куба

З'єднання таблиць ієрархії виміру



**Нормалізовані таблиці вимірів LEFT JOIN,
якщо ієрархія незбалансована (пропущені рівні)**

```
SELECT Customer.City,
       COALESCE(Publisher.Name, 'N/A'),
       SUM(Sales.Amount)
  FROM Sales
 INNER JOIN Customer ON Sales.CustomerKey = Customer.Key
 INNER JOIN Product   ON Sales.ProductKey = Product.Key
 LEFT JOIN Publisher  ON Product.PublisherKey = Publisher.Key
 GROUP BY Customer.City, Publisher.Name
```

**Коефіцієнт із таблиці-мосту для М-М зв'язку
ієрархії, якщо у запиті є відповідна таблиця**

```
SELECT Category.Name,
       SUM(Sales.Amount * Bridge.Weight)
  FROM Sales
 INNER JOIN Product  ON Sales.ProductKey = Product.Key
 INNER JOIN Bridge    ON Bridge.ProductKey = Product.Key
 INNER JOIN Author   ON Bridge.AuthorKey = Author.Key
 GROUP BY Category.Name
```

Зважена метрика кількості

Багатозначний вимір дублює рядки таблиці фактів. Агрегатні функції мають враховувати ваговий коефіцієнт, щоб загальна сума була однаковою на всіх рівнях презентації даних

Fact
P1
P2

Bridge Table

Product	Author
P1	A1
P1	A2
P2	A1
P2	A3

Select Query

Author	Count
A1	2
A2	1
A3	1

Product	Author	Weight
P1	A1	0.5
P1	A2	0.5
P2	A1	0.7
P2	A3	0.3

Author	Weighted Count
A1	1.2
A2	0.5
A3	0.3

кількість

$$\sum = 2$$

Вибірка деталізованих даних

- Рядки таблиці фактів на первинному рівні грануляції: всі метрики, виміри та їхні атрибути
- Відбір за вказаними значеннями ключів
- Інтерфейс BI може мати доступ до первинних даних OLTP-систем за збереженим бізнес-ключем
- Необов'язкові виміри, зв'язок багато-до-багатьох та складні ієрархії, задані в додаткових таблицях



```
SELECT Sales.*, Customer.*, Product.*, Publisher.*,
       ( SELECT STRING_AGG(Author.Name)
           FROM Authors
           JOIN Bridge ON Bridge.AuthorKey = Author.Key
           WHERE Bridge.ProductKey = Product.Key )
    FROM Sales
    INNER JOIN Customer ON Sales.CustomerKey = Customer.Key
    INNER JOIN Product   ON Sales.ProductKey = Product.Key
    LEFT JOIN Publisher  ON Product.PublisherKey = Publisher.Key
   WHERE Sales.CustomerKey IN ('c1', 'c2') AND
         Sales.ProductKey = 'p1'
```

Вибірка із факту без метрик

Для всіх значень вимірів
встановлюється наявність факту

SELECT

Year.Name,
Student.Name,
Course.Name,

EXISTS (

SELECT * FROM Fact
WHERE

(YearKey, StudentKey, CourseKey) =
(Year.Key, Student.Key, Course.Key))

FROM Year, Student, Course

		BPM	DW	ADB
2012	John	X	X	X
	Mary			
	Pete			
	Patrick	X	X	
2013	Patrick			X
	Jane		X	
	Pete			X

Для кожного значення виміру
обираються факти за квантором

SELECT

Student.Name,
Student.Key = ANY

(SELECT StudentKey

FROM Fact

WHERE CourseKey = 'DW'),
Student.Key = ALL

(SELECT StudentKey

FROM Fact

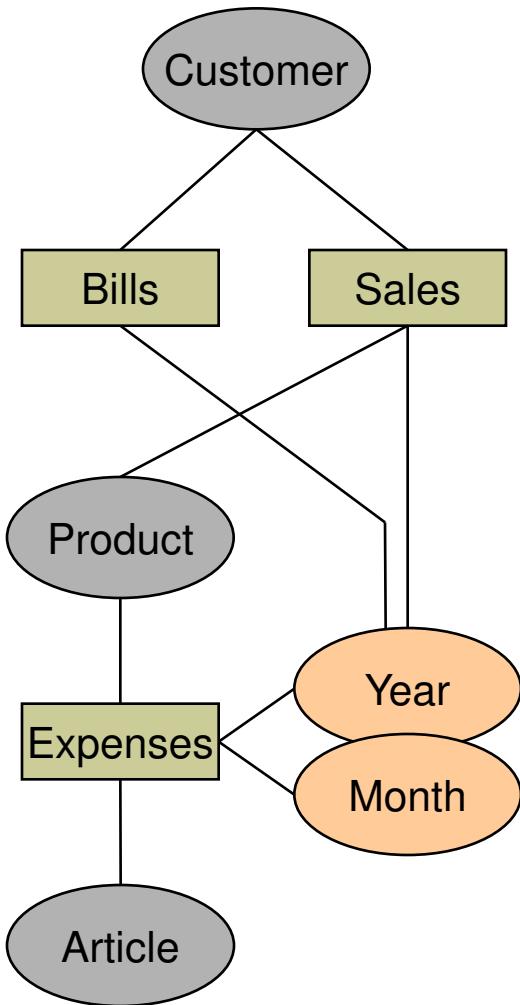
RIGHT JOIN Course

ON (StudentKey, CourseKey) =
(Student.Key, Course.Key))

FROM Student

	Any	All
John	X	X
Mary		
Pete		
Patrick	X	X
Jane	X	

Запити на об'єднання фактів



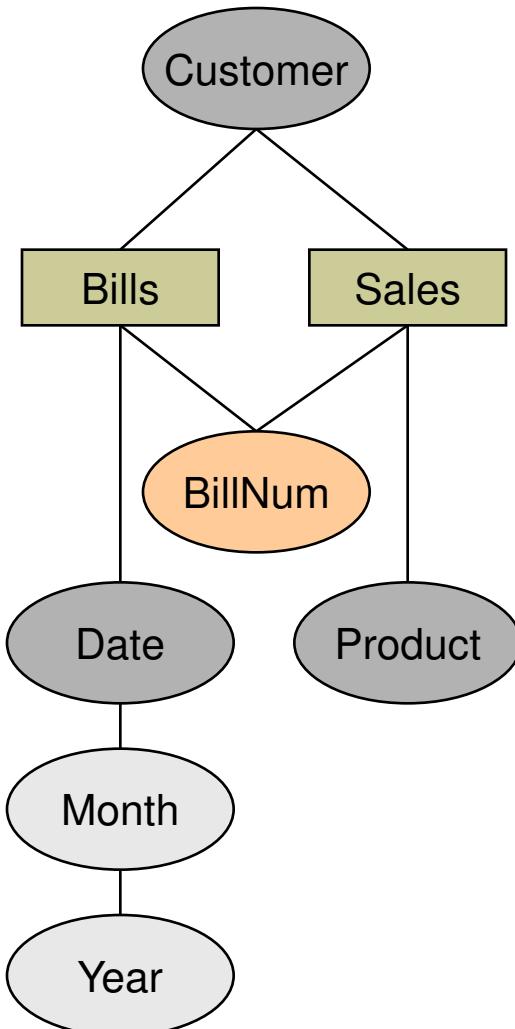
Об'єднання фактів UNION ALL
виконується у підзапиті, який формує
КОЛОНКИ СПІЛЬНИХ ВИМІРІВ та метрик

```
SELECT Un.Year, Un.Month,
SUM(Un.Total), SUM(Un.Amount)
FROM (
    SELECT CustomerKey, Year, Month, Total, NULL
    FROM Bills
    UNION ALL
    SELECT CustomerKey, Year, Month, NULL, Amount
    FROM Sales
) AS Un
GROUP BY Un.Year, Un.Month
```

Використання функції COALESCE,
щоб врахувати відсутні значення

```
SELECT SUM(COALESCE(Un.Amount, 0) -
COALESCE(Un.Total, 0))
FROM ( ... ) AS Un INNER JOIN Customer
ON Un.CustomerKey = Customer.Key
```

З'єднання із фактом відстеження



Можливо застосування INNER JOIN, якщо вироджений вимір реалізує зв'язок 1-М, перевірку рівності значень можна відкласти з метою оптимізації

```
SELECT Date.Month, SUM(Sales.Amount)
  FILTER (WHERE Bills.BillNum = Sales.BillNum)
FROM Bills INNER JOIN Sales
  ON Bills.CustomerKey = Sales.CustomerKey
INNER JOIN Date
  ON Date.Key = Bills.DateKey
 WHERE Date.Year IN (2012, 2013, 2014)
 GROUP BY Date.Month
```

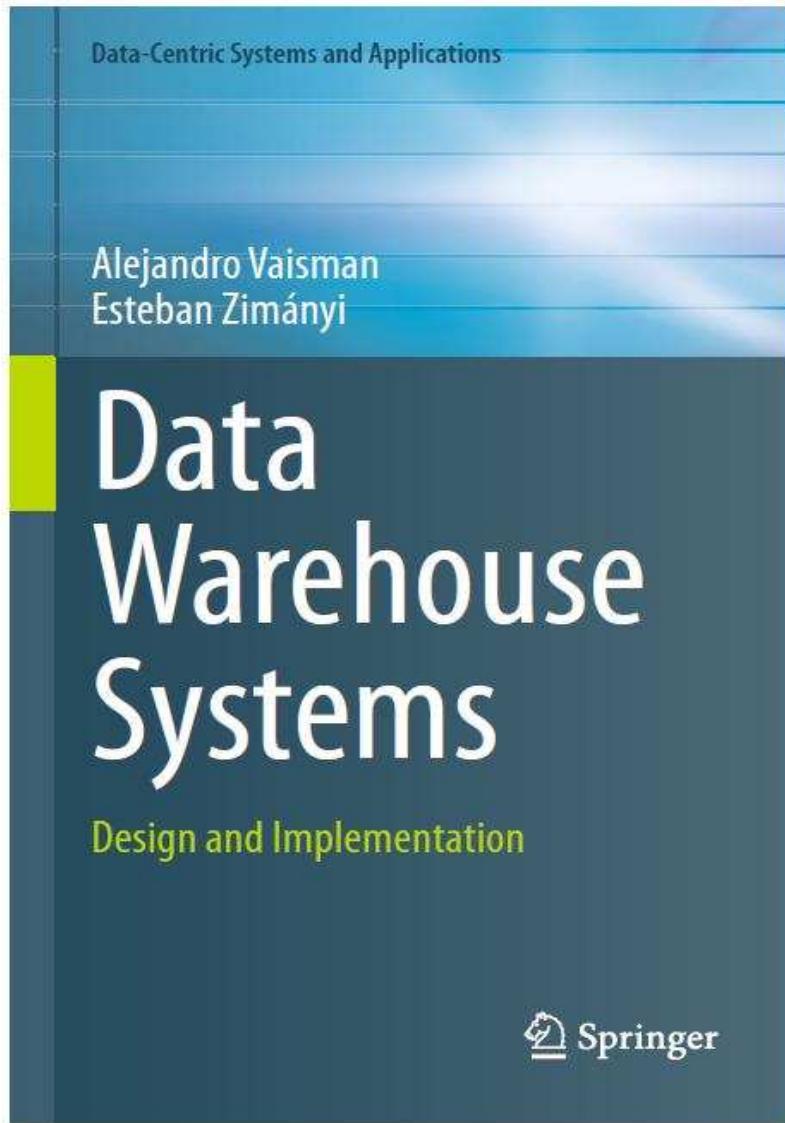
Усунення дублікатів DISTINCT при виведенні метрик факту відстеження

```
SELECT Bills.Date,
  COUNT(DISTINCT Bills.BillNum),
  SUM(DISTINCT Bills.Total)
FROM Bills, Sales ...
```

Таблична форма зберігання кубів

- Вибірка виконується “вертикально” з таблиці фактів з фільтрацією за ключами, отриманими з таблиць вимірів
 - Враховується структура ієрархії та зв’язків між рівнями
 - При з’єднанні таблиць рядки можуть дублюватися, що приводить до некоректного обчислення метрик
 - Скерувати запит до факту-агрегату або матеріалізованому представленню, якщо вони є на відповідному рівні
 - Агрегація значень метрик відбувається в пам’яті, очікуваний об’єм вибірки з кубу є невеликим
-
- Трансляція багатовимірних запитів користувача в реляційну модель кубу потребує знань:
 - структурної схеми кубу
 - логічної схеми бази даних
 - відповідності елементів кубу таблицям та колонкам
 - технік оптимізації сховища даних

Дякую за увагу!



The Data Warehouse Toolkit

Third Edition

The Definitive Guide
to Dimensional
Modeling

Ralph Kimball
Margy Ross

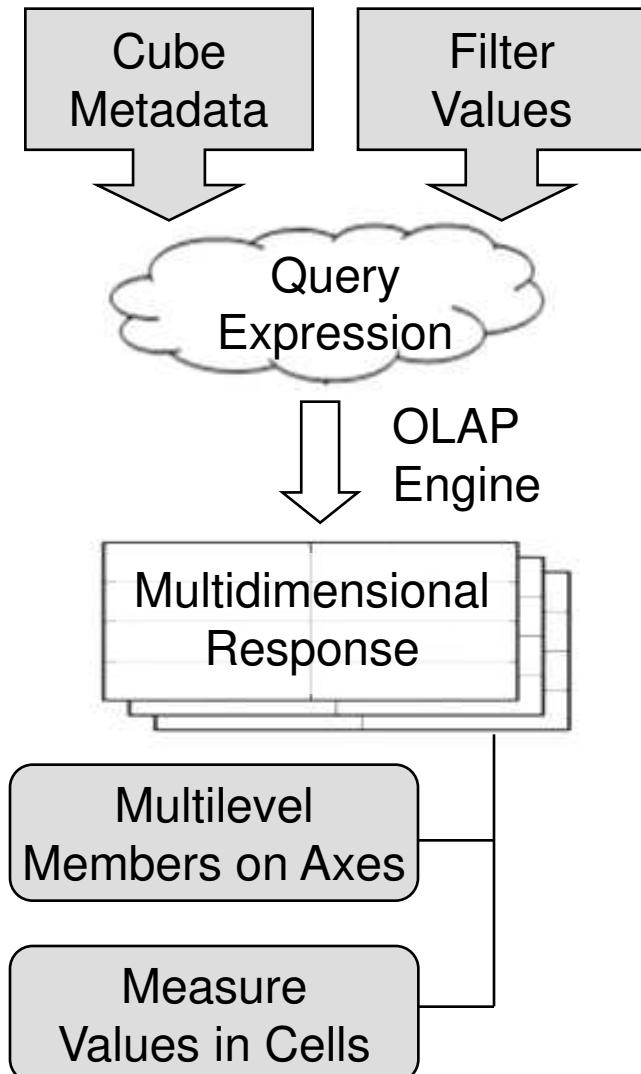


СХОВИЩА ДАНИХ: Лекція №9

НУ “Львівська Політехніка”, кафедра ПЗ

Мова запитів до багатовимірних
структур даних

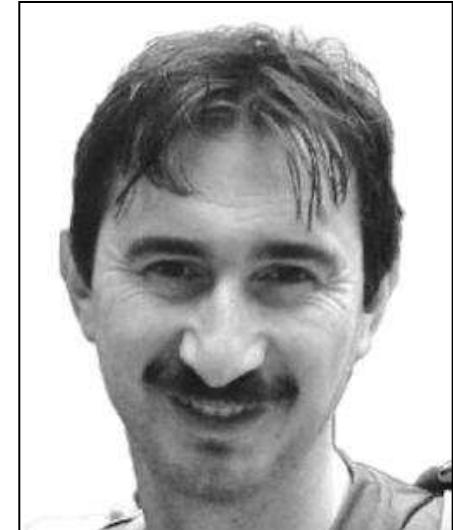
Витяг багатовимірних даних



- Користувачем визначаються множини елементів вимірів та назви метрик, які будуть включені до результату витягу
- Встановлюються умови фільтрації, зрізи по осіах, вирази обчислення значень, сортування та ін.
- Опрацьовується ієрархія вимірів, метрики агрегуються автоматично відповідно до їхніх типів
- Будується ефективний план, враховуючи індекси, обчислені агрегати, поточний кеш, можливості паралельного виконання та ін.
- Дані читаються з рівня зберігання відповідно до технології
- Структурований результат надається як багатовимірний куб

Мова від Моши Пасуманського

- **Mosha Pasumansky (born 1970)**
received a M.Sc. in Computer Science
from the University of Washington
- Co-inventor of the Multidimensional
Expressions Language (MDX)
- From 1996 was a member of the OLAP
development team in Microsoft, one of
the architects of the Analysis Services.
Became a representative from Microsoft
in XMLA Council
- From 2011 to 2021 worked at Google
on the Dremel and BigQuery systems
- “Fast Track to MDX, 2nd Edition” (2005)



Запити MDX: семантична модель кубу

- **MultiDimensional eXpressions** is a language for querying and manipulating the data stored in cubes
- Specification was first introduced in 1997 by **Microsoft**
- Became de-facto the standard supporting by **Oracle**,
SAP HANA, **IBM Cognos**, **SAS Server**, **Hitachi Ventara** and other analytical platforms
- Describes what the data to fetch and how to format the results in terms of multidimensional modeling
- Works over cubes, dimensions, hierarchies, and members at the instance level
- Supports definition of calculated members, named sets, scoped assignments, key performance indicators
- Allows to add business logic to the cubes, to define custom roll-ups and actions, to define security settings

Запити DAX: формули табличної моделі

- **Data Analysis Expressions** is the native formula and query language for the **Microsoft** data analytical products based on tabular models
- First released in 2009 as an evolution of the MDX to be easy to learn
- Includes some of the **Excel** formulas and additional functions that are designed to work with relational data and perform dynamic aggregation
- Defines custom calculations for the parts of OLAP model, filter expressions in role-based security in tabular models

```
DEFINE MEASURE 'Internet Sales'[Internet Total Sales] = SUM('Internet Sales'[Sales Amount])
EVALUATE SUMMARIZECOLUMNS
(
    'Date'[Calendar Year],
    TREATAS({2013, 2014}, 'Date'[Calendar Year]),
    "Total Sales", [Internet Total Sales],
    "Combined Years Total Sales",
    CALCULATE([Internet Total Sales], ALLSELECTED('Date'[Calendar Year]))
)
ORDER BY [Calendar Year]
```

Messages	Results
Date[Calendar Y...	[Total Sales]
2013	16351550.34
2014	45694.72

Date[Calendar Y...	[Total Sales]	[Combined Years...
2013	16351550.34	16397245.06
2014	45694.72	16397245.06

Специфікація XMLA: дані для аналітики

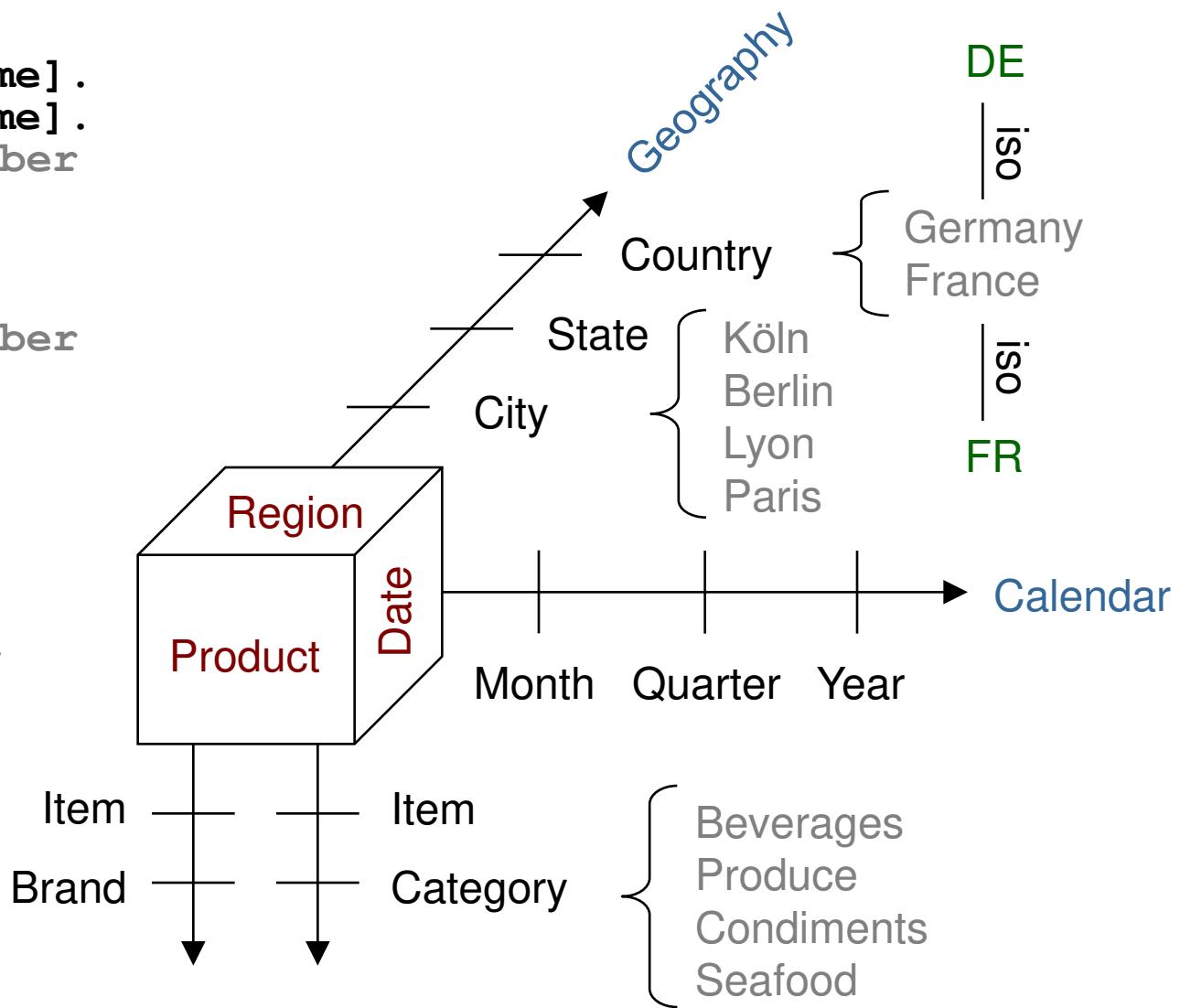
- **XML for Analysis** is an industry standard for data access in analytical systems, such as online analytical processing and data mining
- Specification was first proposed by **Microsoft** in 2000 as a successor for OLE DB, and now it is maintained by XMLA Council with **Hyperion and SAS Institute**
- Supports the discovery and manipulation of the data presented in both tabular and multidimensional modes
- Based on SOAP messaging protocol to communicate clients with the data analysis services
- Implements **Tabular Model Scripting Language** and **Multidimensional Expressions** making OLAP cube accessible to end-user applications

Терміни багатовимірних виразів

- **Dimension** is organized by the named **Levels** in own named **Hierarchy**, they are assumed to be mutually independent [D].[H].[L]
- **Member** represents the axis point of the cube and is defined by a unique name within its dimension, level and parent members [A].[1].[2].[3]
- **Set** is ordered collection of members with the same dimensionality or hierachality { [A].[1], [A].[2] }
- **Tuple** contains list of the members and sets from the differenet dimensions ([A], { [1], [2] })
- **Cell** with the data is pointed by the tuple at the corresponding level of aggregation
- **Property** is the descriptive attribute of the member, and its value can be utilized in a query
- **Measure** acts like dimension in a query context

Вираз множини елементів виміру

- [Dimension_Name].
[Hierarchy_Name].
DefaultMember
AllMembers
- [Level_Name].
CurrentMember
Members
- [Member_1].
[Member_2].
[Member_N].
PrevMember
FirstChild
Parent
Children
Hierarchy
- Properties ("Name")



Кортеж для ідентифікації комірок

1

(Date.Quarter.Q1,
Product.Category.Beverages,
Region.City.Paris)
= 21

2

(Date.Quarter.Q1,
Region.City.Paris)
= 21 + 10 + 18 + 35

3

(Date.Quarter.Q1,
Product.Category.Beverages,
{ Region.Country.France,
Region.Country.Germany })
= 21 + 12 + 33 + 24

4

(Date.Quarter.Q1)
(Product.Category.Seafood)
(Region.City.Paris)
= all visible cells

		Köln	24	18	28	14	
	Berlin	33	25	23	25	14	
	Lyon	12	20	24	33	25	
	Paris	21	10	18	35	33	18
Q1	21	10	18	35	35	23	
Q2	27	14	11	30	30	20	18
Q3	26	12	35	32	32	33	
Q4	14	20	47	31	31	10	
	Beverages	Produce	Condiments	Seafood			

Структура запиту на вибірку

```
WITH MEMBER [Calculated Value]
      Measures . [Avg Bill]
      AS ( Measures . [Sales Amount] /
            Measures . [Sales Count] )

SELECT [Returns Set]
      { Measures . Members ,
        Measures . [Avg Bill] }

ON COLUMNS ,
{ Date . Calendar . [1996] . [Q1] ,
  Date . Calendar . [1996] . [Q2] }

ON ROWS [Axis Specification]

FROM Sales

WHERE ( { Region . Country . France ,
          Region . City . Berlin } )
```

- Query defines several result axes, the first five names are predefined
- Requesting result with more than one dimension makes each cell appearing in the associated axis
- Square brackets around the object identifier are optional if it is not one of the reserved words
- Curly brackets are representing a set of members from the same dimension
- Slicer determines which members of the chosen dimensions will be used as extraction filter

Призначення клауз та функцій

- DRILLTHROUGH
 - WITH MEMBER | SET *name*
 - SELECT NON EMPTY *set* DIMENSION PROPERTIES ON AXIS (*int*)
 - FROM *cube* | (...) CELL PROPERTIES
 - WHERE (..., { ... })
-
-
- ASCENDANTS (*set*)
 - DESCENDANTS (*set*)
 - EXCEPT (*set1*, *set2*)
 - CROSSJOIN (*set1*, *set2*)
 - FILTER (*set*, *expression*)
 - ORDER (*set*, *measure*)
 - HEAD (*set*, *count*)
- деталізація даних вибірки
 - вирази, функції для розрахунку окремих значень або множин
 - специфікація осей запиту
 - пропуск елементів без даних
 - вибір вказаних атрибутів
 - порядковий номер осі
 - куб або вкладений підзапит
 - вибір властивостей комірок
 - кортеж множин зрізу по осях
-
-
- елементи в напрямку узагальнення ієархії, всі вкладені в набір
 - перетин множин елементів, комбінування кожного з кожним
 - відбір елементів за умовою
 - впорядкування за значенням
 - відбір вказаного числа елементів

■ Витяг всіх метрик по 1 виміру на встановленому рівні

```
SELECT [Measures].MEMBERS ON COLUMNS,  
       [Customer].[Country].MEMBERS ON ROWS  
FROM   Sales
```

	Unit Price	Quantity	Discount	Sales Amount	Freight	Sales Count
Austria	\$84.77	4,644	21.71%	\$115,328.31	\$6,827.10	114
Belgium	\$64.65	1,242	9.72%	\$30,505.06	\$1,179.53	49
Denmark	\$70.28	1,156	17.94%	\$32,428.94	\$1,377.75	45

■ Витяг 1 метрики по 2 вимірах із зрізом по категорії

```
SELECT [Order Date].Year.MEMBERS ON COLUMNS,  
       Customer.Country.MEMBERS ON ROWS  
FROM   Sales  
WHERE  (Measures.[Sales Amount], Product.Category.[Beverages])
```

	All	1996	1997	1998
Austria	\$115,328.31	\$24,467.52	\$55,759.04	\$35101.7502
Belgium	\$30,505.06	\$5,865.10	\$9,075.48	\$15,564.48
Denmark	\$32,428.93	\$2,952.40	\$25,192.53	\$4,284.00

■ Витяг вкладених елементів з певного рівня ієрархії

```
SELECT [Order Date].Year.MEMBERS ON COLUMNS,  
       DESCENDANTS(Customer.Germany, Customer.City) ON ROWS  
FROM   Sales  
WHERE  Measures.[Sales Amount]
```

	All	1996	1997	1998
Mannheim	\$2,381.80	\$1,545.70	\$1,079.80	\$1,302.00
Stuttgart	\$8,705.23	\$2,956.60	\$4,262.83	\$1,485.80
München	\$26,656.56	\$9,748.04	\$11,829.78	\$5,078.74

■ Витяг елементів по рівням ієрархії відносно заданого

```
SELECT Measures.[Sales Amount] ON COLUMNS,  
       ASCENDANTS(Customer.Geography.[Nantes]) ON ROWS  
FROM   Sales
```

	Sales Amount
Nantes	\$4,720.86
Loire-Atlantique	\$4,720.86
France	\$77,056.01
Europe	\$683,523.76
All Customers	\$1,145,155.86

City
State
Country
↓Continent

■ Фільтр елементів по зрізу формує вісь результатів

```
SELECT Product.Category.MEMBERS ON COLUMNS,  
       FILTER(Customer.City.MEMBERS,  
              (Measures.[Sales Amount], [Order Date].Calendar.[1997])  
              > 25000) ON ROWS  
FROM   Sales  
WHERE  (Measures.[Net Sales Growth],  
        [Order Date].Calendar.[1997])
```

	Beverages	Condiments	Confections	Dairy Products
Montréal	\$9,142.78	\$2,359.90	\$213.93	\$3,609.16

■ Елементи вісі впорядковані за значенням метрики

```
SELECT Measures.MEMBERS ON COLUMNS,  
       HEAD(ORDER(Customer.Geography.City.MEMBERS,  
                  Measures.[Sales Amount], BDESC), 3) ON ROWS  
FROM   Sales
```

	Unit Price	Quantity	Discount	Sales Amount	Freight	Sales Count
Cunewalde	\$101.46	3,616	21.40%	\$103,597.43	\$4,999.77	77
Boise	\$90.90	4,809	32.41%	\$102,253.85	\$6,570.58	113
Graz	\$88.00	4,045	23.57%	\$93,349.45	\$5,725.79	92

■ Обчислення за формулою значення нової метрики

WITH MEMBER Measures.Profit AS

$$(\text{Measures.[Sales Amount]} - \text{Measures.[Freight]}) / (\text{Measures.[Sales Amount]})$$

SELECT { [Sales Amount], Freight, Profit } ON COLUMNS,
Customer.Country ON ROWS

	Sales Amount	Freight	Profit
Austria	\$115,328.31	\$6,827.10	94.08
Belgium	\$30,505.06	\$1,179.53	96.13

■ Формування набору елементів за результатом функції

WITH SET TopFiveProducts AS

$$\text{TOPCOUNT} (\text{Product.Categories.Product.MEMBERS}, 5, \text{Measures.[Sales Amount]})$$

SELECT { Measures.Quantity, Measures.Discount, Measures.[Sales Amount] }
ON COLUMNS, TopFiveProducts ON ROWS

	Quantity	Discount	Sales Amount
Côte de Blaye	623	4.78%	\$141,396.74
Raclette Courdavault	1,369	3.96%	\$65,658.45
Thüringer Rostbratwurst	596	6.21%	\$63,657.02
Tarte au sucre	1,068	5.53%	\$46,643.97

■ Підзапит виконує зріз по 2 категоріях одного виміру

```
SELECT Measures.[Sales Amount] ON COLUMNS,  
       [Order Date].Calendar.Quarter.MEMBERS ON ROWS  
FROM   ( SELECT { Product.Category.Beverages,  
                  Product.Category.Condiments }  
        FROM   Sales )  

```

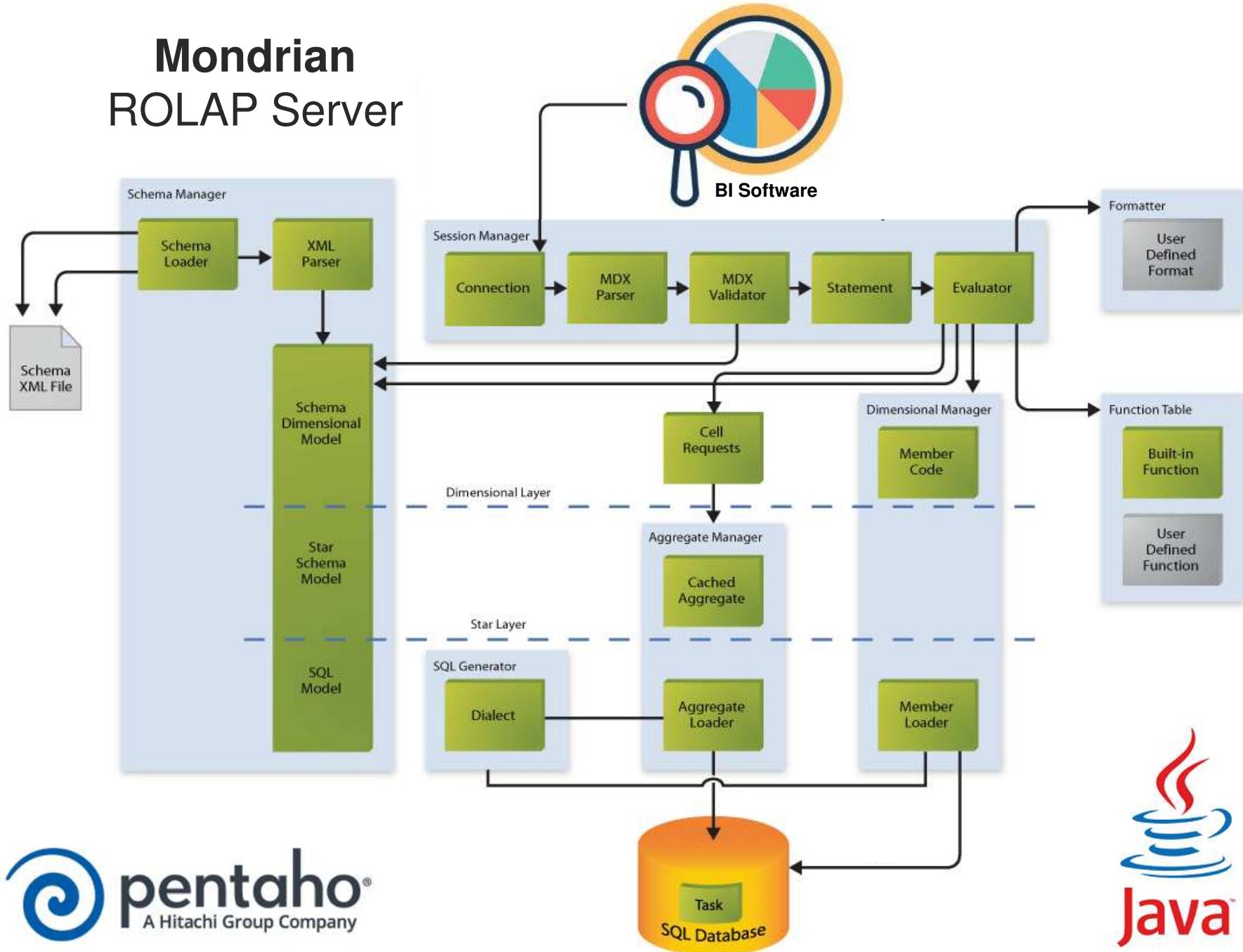
```
SELECT Measures.[Sales Amount] ON COLUMNS,  
       [Order Date].Calendar.Quarter.MEMBERS ON ROWS  
FROM   Sales  
WHERE  { Product.Category.Beverages, Product.Category.Condiments }
```

■ Перетин вимірів створює групи по обраній вісі

```
SELECT Product.Category.MEMBERS ON COLUMNS,  
       Customer.Country.MEMBERS *  
       [Order Date].Calendar.Quarter.MEMBERS ON ROWS  
FROM   Sales WHERE Measures.[Sales Amount]
```

		Beverages	Condiments	Confections
Austria	Q3 1996	\$708.80	\$884.00	\$625.50
Austria	Q4 1996	\$12,955.60	\$703.60	\$36.00
Austria	Q1 1997	\$2,610.51	\$3,097.50	\$1,505.22
Austria	Q2 1997	\$1,287.50	\$1,390.95	\$3,159.00

Mondrian ROLAP Server



Schema Workbench

- Створення файлу метаданих
- Опис кубу і зв'язок з таблицями
- Виконання MDX-запитів в режимі ROLAP при з'єднанні з RDBMS

The screenshot shows the Schema Workbench interface. On the left is the 'Schema' browser with a tree view of objects like Sales, Data, Theme, and Time. In the center is the 'Level for Hierarchy' editor showing attribute values for 'Hour'. At the bottom is the 'MDX Query - connected to MySQL' editor displaying an MDX query and its results.

```
SELECT [Дата].[Рік].[2024]
ON COLUMNS
FROM
[Sales]
WHERE
[Measures].[Відсоток націнки]
```

```
Axis #0:
{[Measures].[Відсоток націнки]}
Axis #1:
{[Дата].[2024]}
Row #0: 49,261
```

La Azada

- Підключення до серверу в режимі ROLAP із схемою або XMLA
- Графічна і таблична візуалізація
- Операції презентації куба, запити

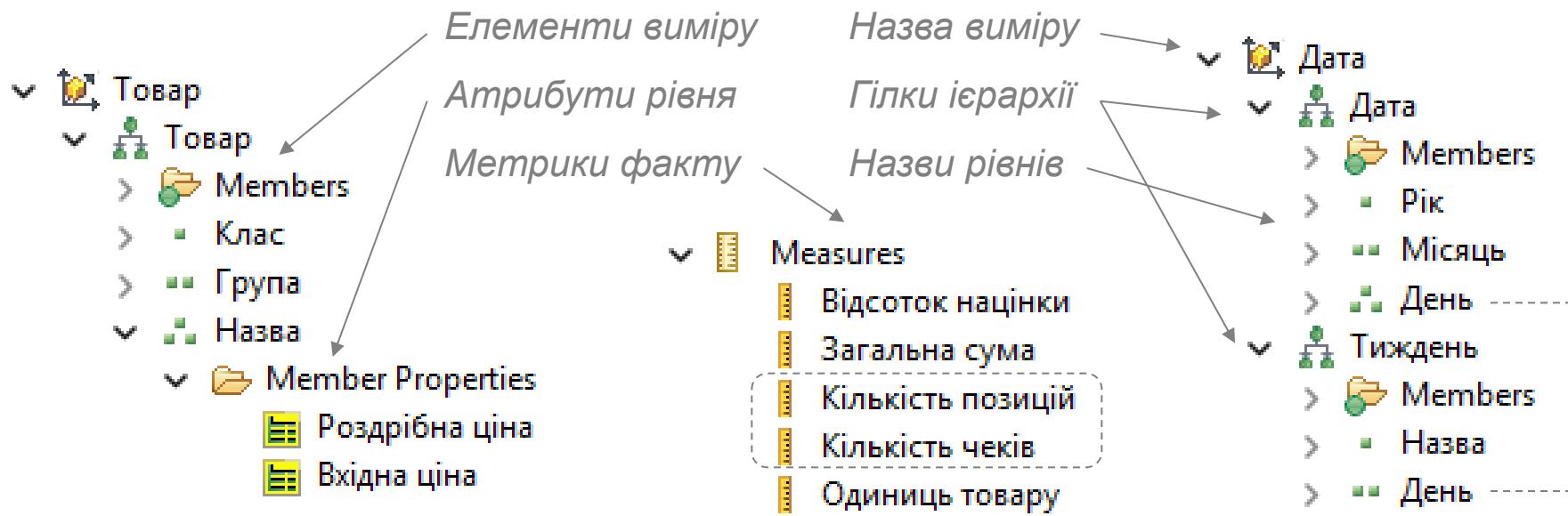
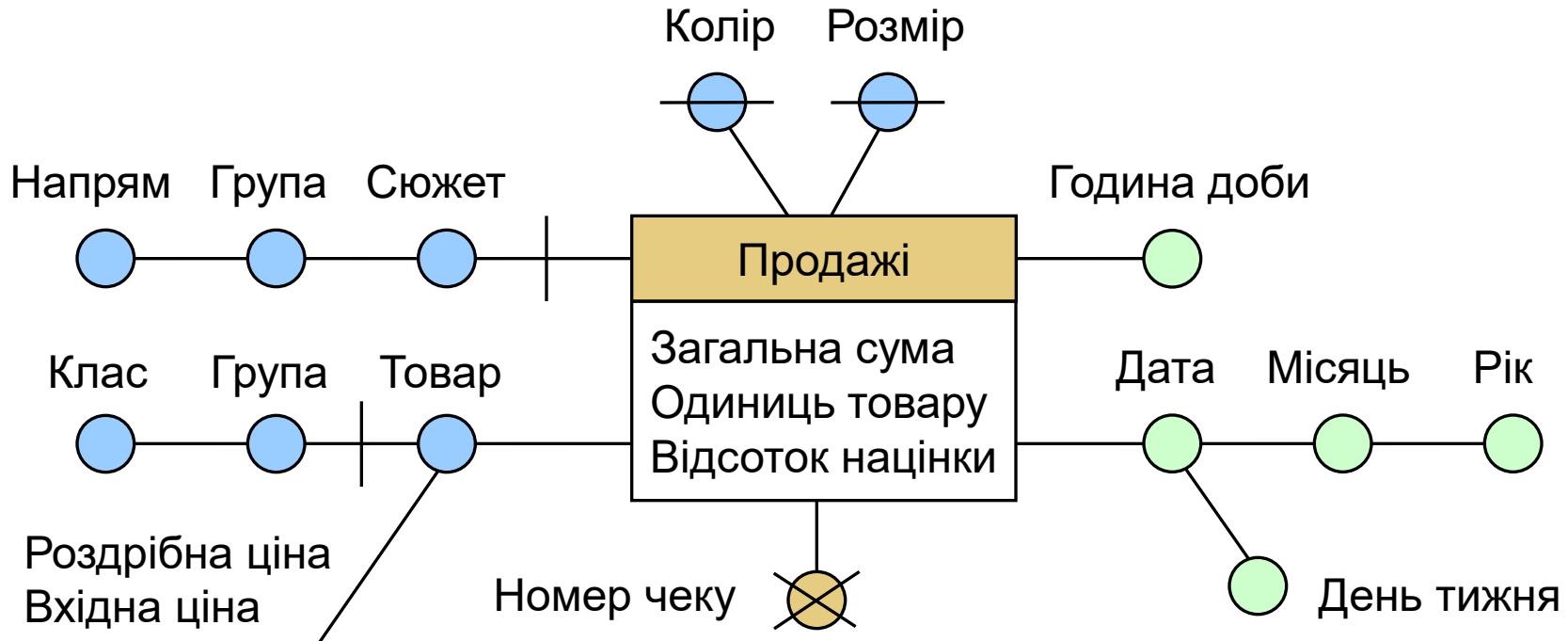
The screenshot shows the La Azada interface. It features a hierarchical tree view of a cube structure, a pivot table showing sales data, and a query editor at the bottom.

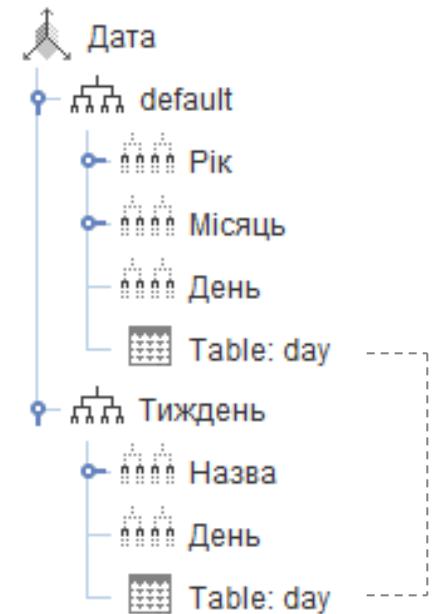
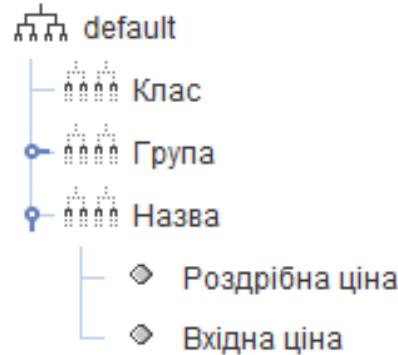
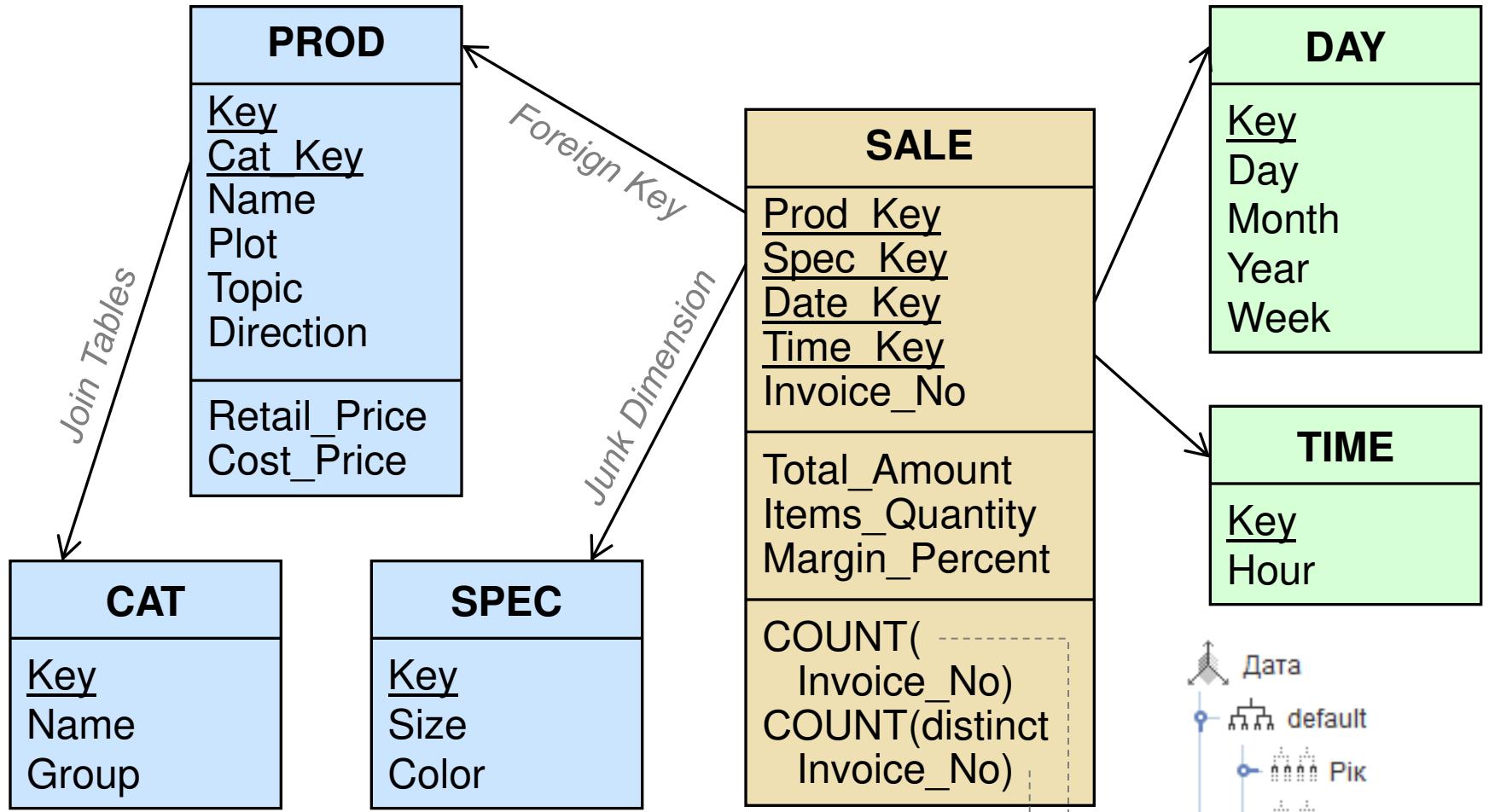
Pivot Table:

Кількість чеків		Дата	Дата	Дата
Напрям	Група	2023	2024	2025
+Анімація		238	1 104	173
+Культура		266	1 808	441
+Кіно		133	555	93
+Музика		213	1 158	241
-Спорт			3	2
Спорт	+Клуби			1
Спорт	+Спортсмени			2
+Ігри		55	221	44

Query Editor:

```
SELECT
(Hierarchize(
{[Дата].[Рік].Members})
ON COLUMNS,
(Hierarchize(
{[Тема].[Напрям].Members,
[Тема].[Спорт].Children})
ON ROWS
FROM
[Sales]
WHERE
{[Measures].[Кількість чеків]}
```





Аналітика кількості проданого товару по місяцях року в розрізі теми

Table Drill-through Chart TreeMap

Дата	Всі	Тема					
		Анімація	Культура	Кіно	Музика	Спорт	Ігри
Jan/2024	451	124	206	47	59	15	
Feb/2024	558	163	237	55	78	1	
Mar/2024	815	228	269	116	167	35	
Apr/2024	596	148	241	58	134	15	
May/2024	584	138	250	46	133	17	
Jun/2024	608	136	248	64	141	19	
Jul/2024	546	127	207	51	142	1	
Aug/2024	1 131	156	690	77	188	1	
Sep/2024	749	112	412	54	142	28	
Oct/2024	588	92	330	52	104	10	
Nov/2024	506	65	293	49	73	6	
Dec/2024	1 073	167	574	73	153	41	
2024	8 205	1 656	3 957	742	1 514	3	
						247	

*Query Sql

```
SELECT
    { [Тема].DefaultMember, [Тема].[Напрям].Members }
ON COLUMNS,
    { [Дата].[2024].Children, [Дата].[2024] }
ON ROWS
FROM [Sales]
WHERE [Measures].[Одиниць товару]
```

Аналітика націнки відповідно до об'єму продажів в розрізі категорії товару

Table Drill-through Chart TreeMap

Товар	Measures					
	Відсоток націнки			Загальна сума		
	Дата		Дата		Дата	
Поліграфія	55,191	55,872	55,05	30 410	127 365	98 624
Головні убори	43,173	53,884	52,119	10 990	132 020	26 392
Сумки	41,1	50,499	59,328	7 830	37 305	12 693
Посуд	52,141	52,023	51,915	27 470	149 255	34 490
Ігра	46,379	49,262	49,965	2 540	12 220	2 310
Одяг	48,428	47,638	47,391	276 523	1 645 076	387 296
Атрибутика	45,485	46,913	49,423	88 160	535 136	121 736,5
Ігри	37,9	39,057	38,7	24 690	79 110	11 220

*Query Sql

```
SELECT
    CROSSJOIN(
        { [Measures].[Відсоток націнки], [Measures].[Загальна сума] },
        { [Дата].[Рік].Members }
    )
ON COLUMNS,
    ORDER(
        [Товар].[Клас].Members, [Відсоток націнки], DESC
    )
ON ROWS
FROM [Sales]
```

Аналітика динаміки продажів одягу по місяцях року з виведенням відсотків

Table Drill-through Chart TreeMap

Дата	Measures		
	Загальна сума	Сума одягу	Приріст одягу
Jan/2024	143 715	73 050	-59,39%
Feb/2024	174 513	96 243	31,75%
Mar/2024	274 175	148 590	54,39%
Apr/2024	225 568	142 808	-3,89%
May/2024	217 063	137 878	-3,45%
Jun/2024	238 510	160 020	16,06%
Jul/2024	201 846	137 316	-14,19%
Aug/2024	339 964	211 504	54,03%
Sep/2024	244 246	151 526	-28,36%
Oct/2024	196 764	124 049	-18,13%

*Query Sql

```
WITH MEMBER
    [Measures].[Сума одягу] AS
        ( [Measures].[Загальна сума], [Товар].[Одяг] )
MEMBER [Measures].[Приріст одягу] AS
    [Measures].[Сума одягу] /
    ( [Measures].[Загальна сума], [Товар].[Одяг],
      [Дата].CurrentMember.PrevMember ) - 1, FORMAT_STRING = "Percent"
SELECT
{ [Measures].[Загальна сума], [Measures].[Сума одягу], [Measures].[Приріст одягу] }
ON COLUMNS,
[Дата].[Місяць].Members
ON ROWS
FROM [Sales]
```

Характеристика поліграфічної продукції без продажів за обраний рік

Table Drill-through Chart TreeMap

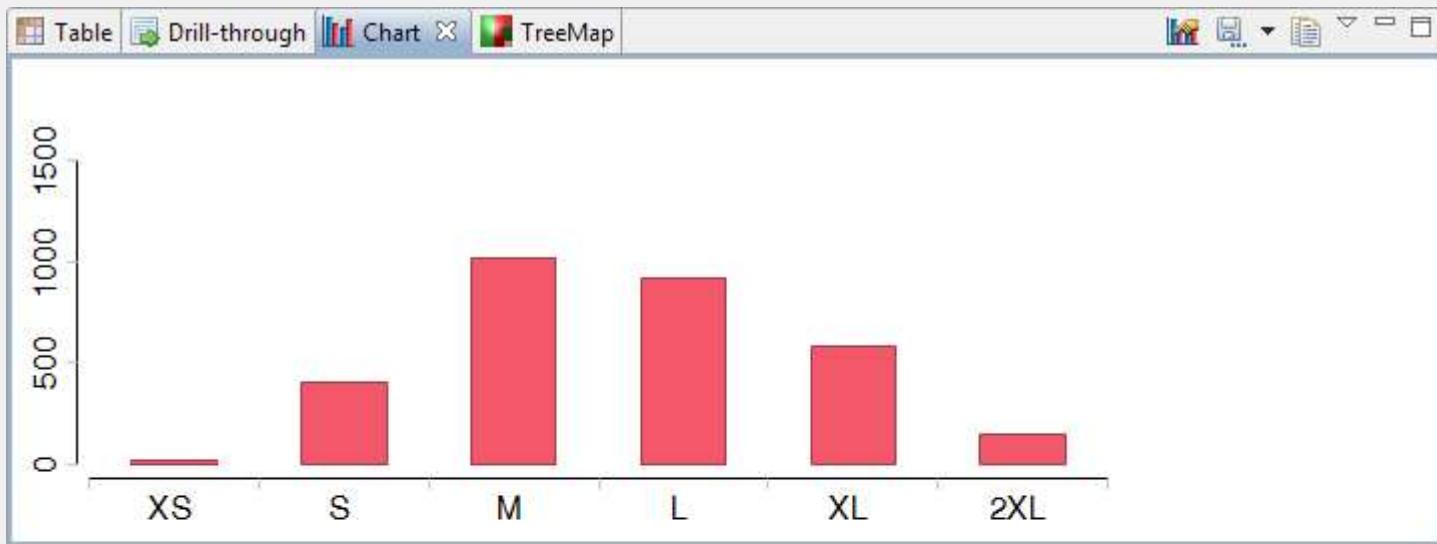
Дата
2024

Товар	Measures
Блокнот на пружині - Kharkiv City, клітинка, 200 стор	590
Блокнот на пружині - Kharkiv City, крапка, 200 стор	590
Скетчбук - Харків, 120 стор	500
Скетчбук на пружині - Kharkiv City, 200 стор	490
Журнал - Наш Street-Art стоп кадр харківських переулков, 2010 р, 24 стор, ВЖИВАНИЙ	190
Комікс Б. Солов'ян, Н. Губський, Р. Аксу - ШЛЯХ А-16, 2023 р, 48 стор	90
Комікс Б.Бендис, В. Скити, Р. Ісанов - Вартові Галактики. Том 1. Імператор Квілл 2020 р,	200

*Query Sql

```
WITH MEMBER
    [Measures].[Роздрібна ціна] AS
        [Товар].CurrentMember.Properties("Роздрібна ціна")
SELECT
    [Measures].[Роздрібна ціна]
ON COLUMNS,
    FILTER(
        DESCENDANTS( { [Товар].[Поліграфія].[Блокноти],
            [Товар].[Поліграфія].[Журнали],
            [Товар].[Поліграфія].[Комікси] } ),
        ISEMPTY( [Measures].[Одиниць товару] )
    )
ON ROWS
FROM [Sales] WHERE [Дата].[Рік].[2024]
```

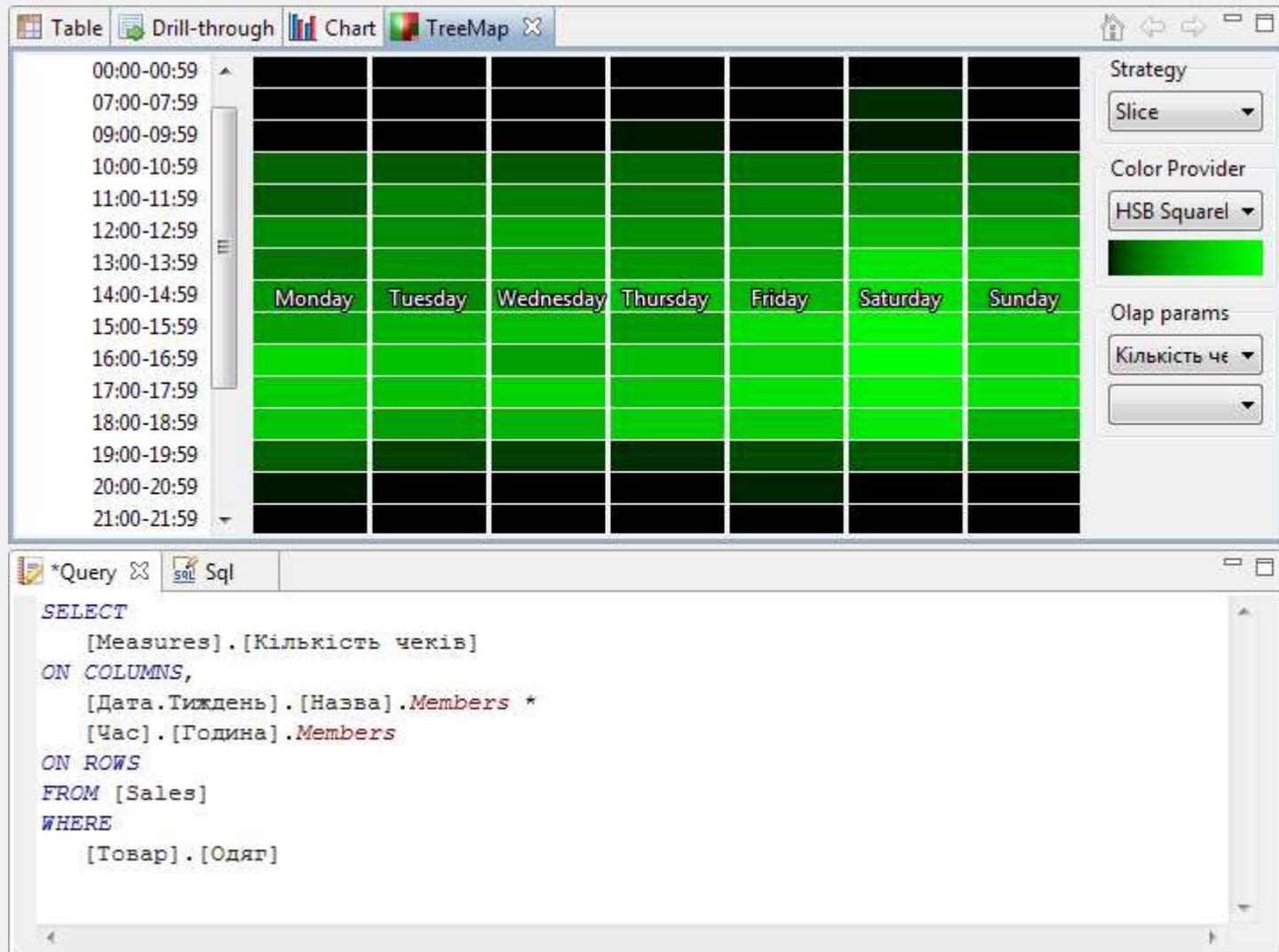
Розподіл проданої кількості одягу по розміру



*Query X SQL

```
SELECT [Measures].[Одиниць товару]
ON COLUMNS,
NON EMPTY EXCEPT(
    [Розмір].[Назва].Members,
    [Розмір].[н/з]
)
ON ROWS
FROM [Sales]
WHERE
    [Товар].[Одяг]
```

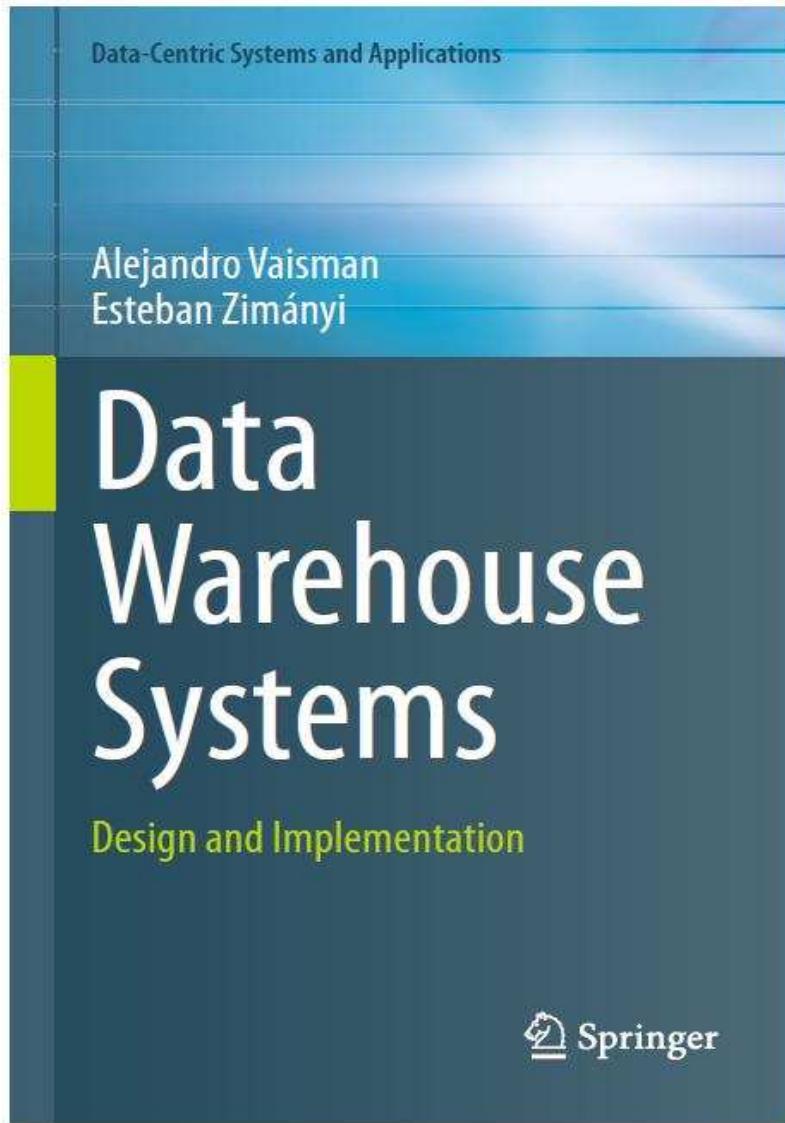
Розподіл кількості покупців по годинах доби та днях тижня



Концепція багатовимірних запитів

- Всі складові кубу мають назви, зберігають порядок при виведенні результатів
- Елементи вимірів ідентифікуються у своїй ієрархії
- Значення формуються на перетині осей, склад яких задається множинами у частині **SELECT**
- Зрізом кубу є кортеж наборів з різних вимірів, що вказується явно у частині **WHERE** або формується як результат функції
- Витяг значень атрибутів, обчислення скалярних виразів та наборів у частині **WITH**
- Запит можливий лише до одного кубу у **FROM**
- Функції над множинами для сортування, фільтрації, обмеження кількості, обходу ієрархії, групування, агрегації, форматування та ін.

Дякую за увагу!



The Data Warehouse Toolkit

Third Edition

The Definitive Guide
to Dimensional
Modeling

Ralph Kimball
Margy Ross



СХОВИЩА ДАНИХ: Лекція №2

НУ “Львівська Політехніка”, кафедра ПЗ

Сховища даних
в корпоративних системах

Дані в процесі бізнес-аналітики

- The Sources of Truth
- Real-Time Data Processing
- Data Transformation and Cleansing
- Data Quality and Governance
- Flexibility and Agility
- Comprehensive Analytics Support
- Cost and Operational Efficiency
- Enhanced Decision-Making



Джерела первинних даних

- Рівні 1-3 для забезпечення технологічних процесів
- Рівні 4-5 для бізнес-аналітики та планування



Корпоративні інформаційні системи накопичення та обробки даних

- **CRM (Customer Relationship Management)**
 - управління взаємовідносинами із споживачами; збір та аналіз інформації про клієнтів, постачальників, партнерів; фіксація та планування маркетингових, комунікативних дій
- **MRP (Manufacturing Resource Planning)**
 - облік ресурсів виробничого підприємства; логістичне та фінансове планування, аналіз потреб, моделювання
- **MES (Manufacturing Execution System)**
 - вирішення завдань синхронізації, координації, аналізу та оптимізації випуску продукції в циклах виробництва
- **SCM (Supply Chain Management)**
 - планування та контроль за потоками інформації у ланцюгу постачання для задоволення потреб клієнтів
- **WMS (Warehouse Management System)**
 - управління складськими та логістичними процесами (облік операцій, планування, запаси, звітність, персонал)
- **PLM (Product Lifecycle Management)**
 - управління інформацією про виріб і пов'язаних з ним процесів протягом життєвого циклу; від проектування і виробництва до зняття з експлуатації

Планування ресурсів: модульний підхід

■ **ERP (Enterprise Resource Planning)**

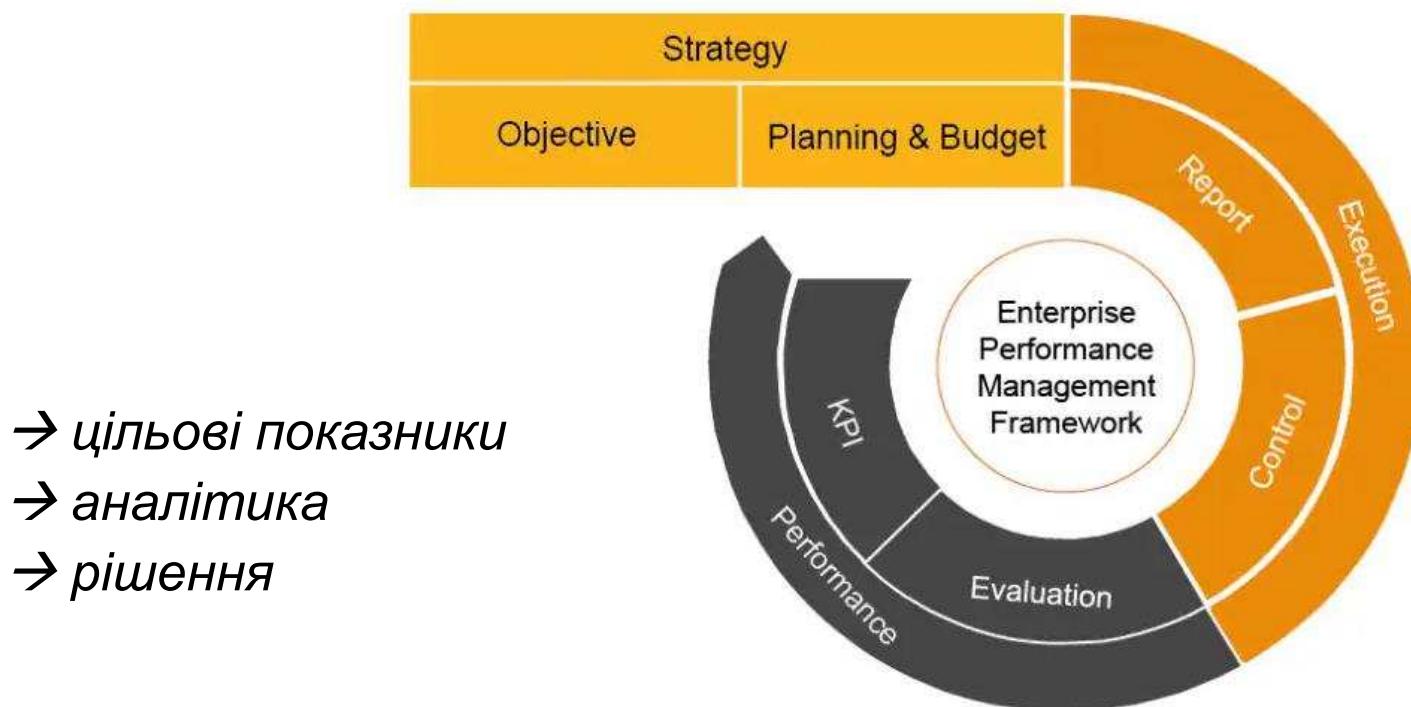
- автоматизація обліку й керування; охоплює ключові процеси діяльності компанії в єдиному інтегрованому середовищі



Планування показників: призначення бізнес-аналітики

■ EPM (Enterprise Performance Management)

- процеси планування, організації виконання, контролю та аналізу, які дозволяють бізнесу визначати цілі та керувати діяльністю задля їхнього досягнення



Типові завдання аналізу даних

□ Аналіз каналів продажу

- Хто з торгових представників краще за всіх продає товари (послуги)?
- Як змінюється ціна на товар (послугу) в різних філіалах компанії?
- Які з партнерів забезпечують найбільший прибуток?
- Які продукти, групи продуктів найкраще продає певний партнер?

Типові завдання аналізу даних

- Аналіз клієнтської бази
 - Які сегменти ринку забезпечують найбільший прибуток?
 - Які клієнти дають найбільший прибуток?
 - Які властивості характерині клієнтам, що забезпечують найбільший прибуток?
- Маркетинг
 - Яка вірогідність відгуку певного сегменту ринку на нову пропозицію?

Типові завдання аналізу даних

□ Аналіз прибутковості

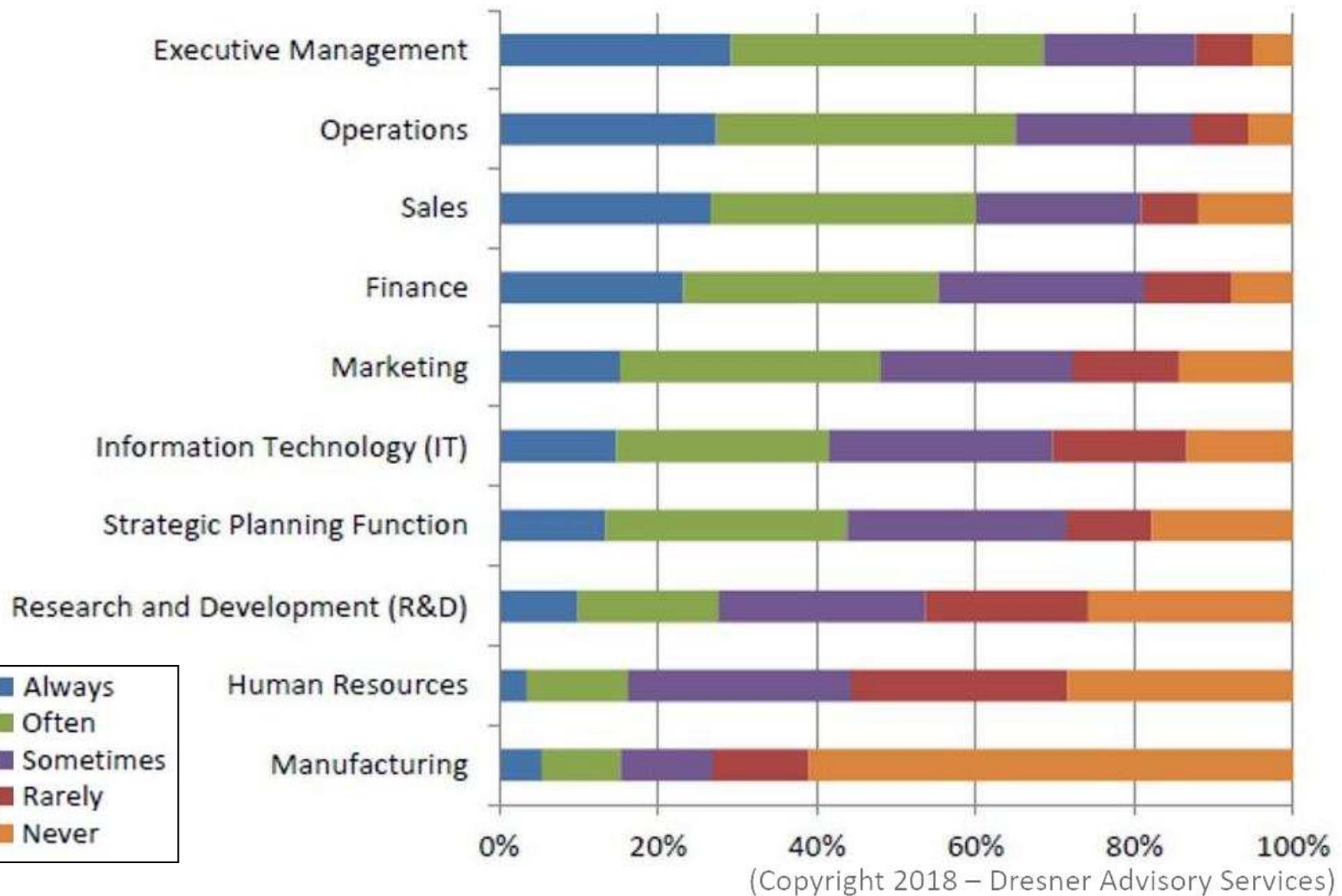
- Які продукти (послуги) дають найбільший прибуток?
- Яка комбінація підрозділів і товарів (послуг) просуває бізнес?
- Які сегменти ринку дають найбільший прибуток?
- Які клієнти забезпечують найбільший прибуток?

Типові завдання аналізу даних

□ Фінансовий аналіз

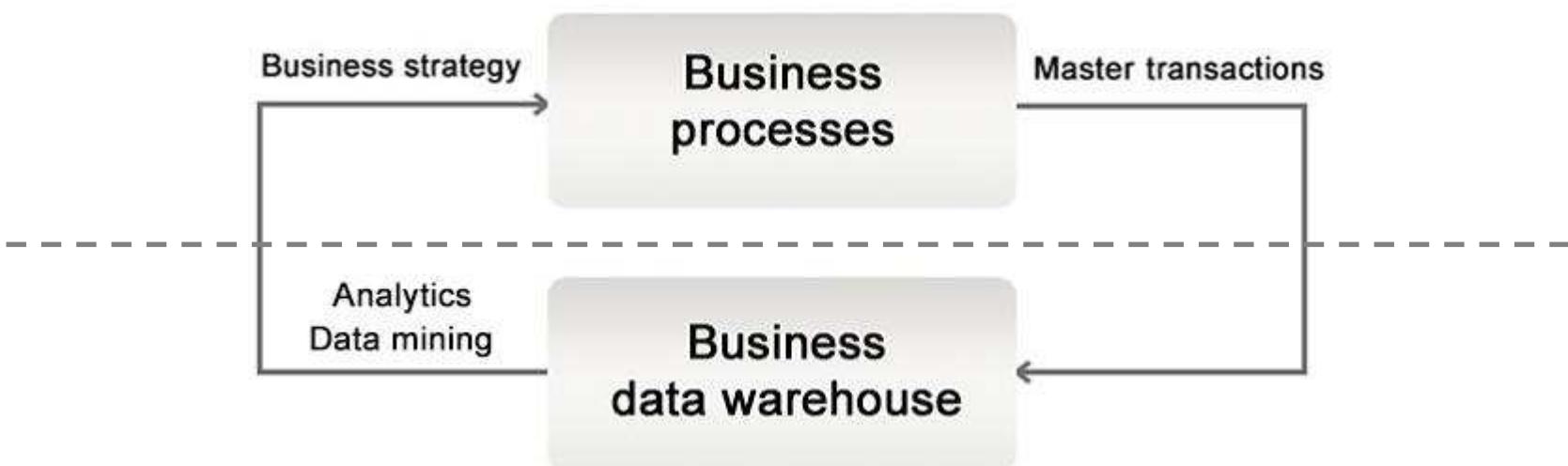
- Наскільки точно різні підрозділи компанії виконують встановлений бюджет?
- Які тенденції витрат за різними підрозділами, статтями бюджету?
- Наскільки вчасно надходять платежі?

Області застосування бізнес-аналітики

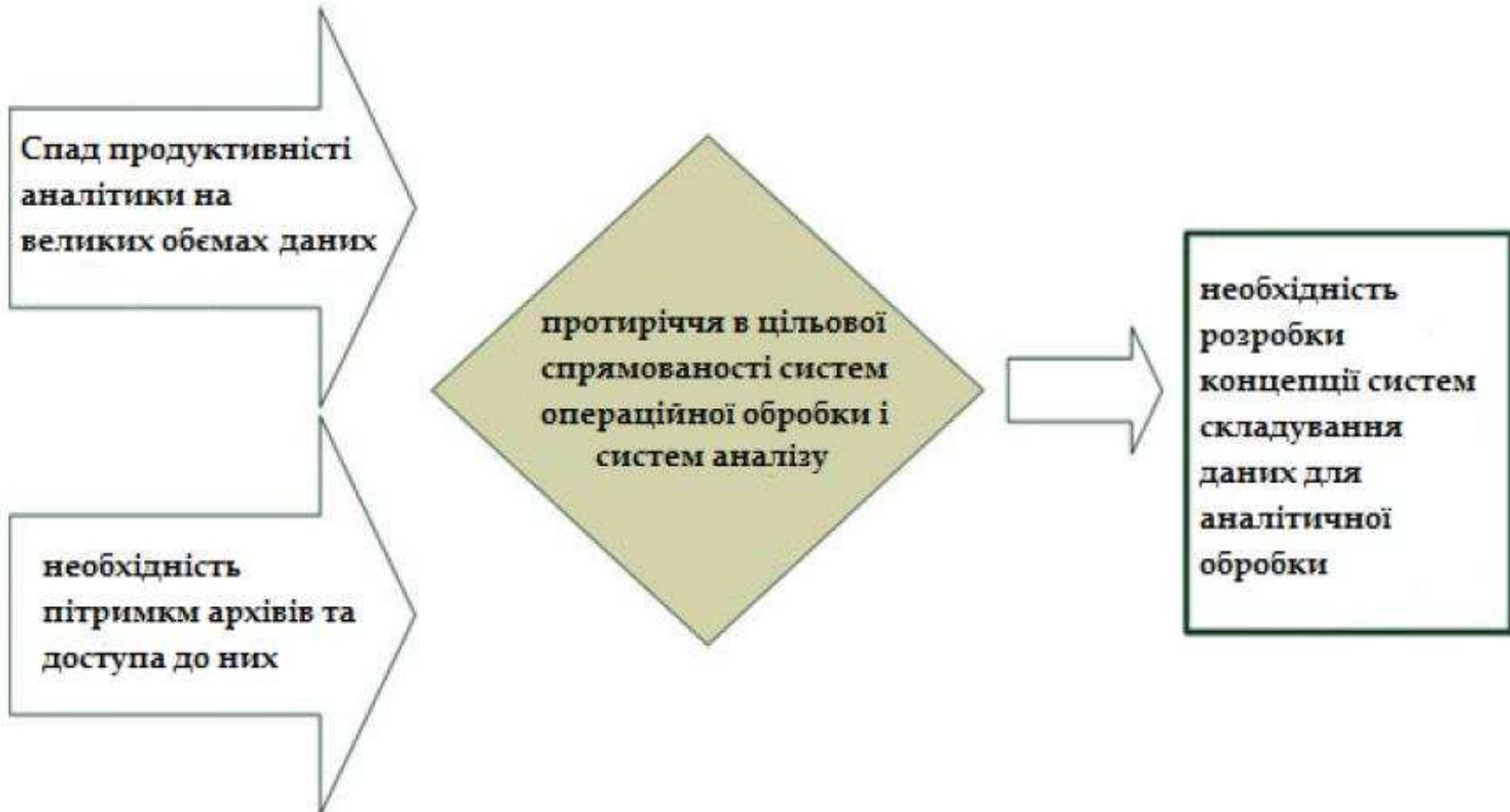


Цільове призначення систем

- **OLTP** – оперативна обробка транзакцій бізнес-процесів
вимоги: постійне оновлення, стабільна швидкість, гарантована цілісність, регламентовані запити, первинні облікові дані
- **OLAP** – підтримка даних для бізнес-аналітики
вимоги: великий обсяг при періодичному оновленні, довільні вибірки, зведені підсумки, консолідовані дані, повні історичні дані



Передумови окремої обробки даних



Порівняння: Performance Requirements

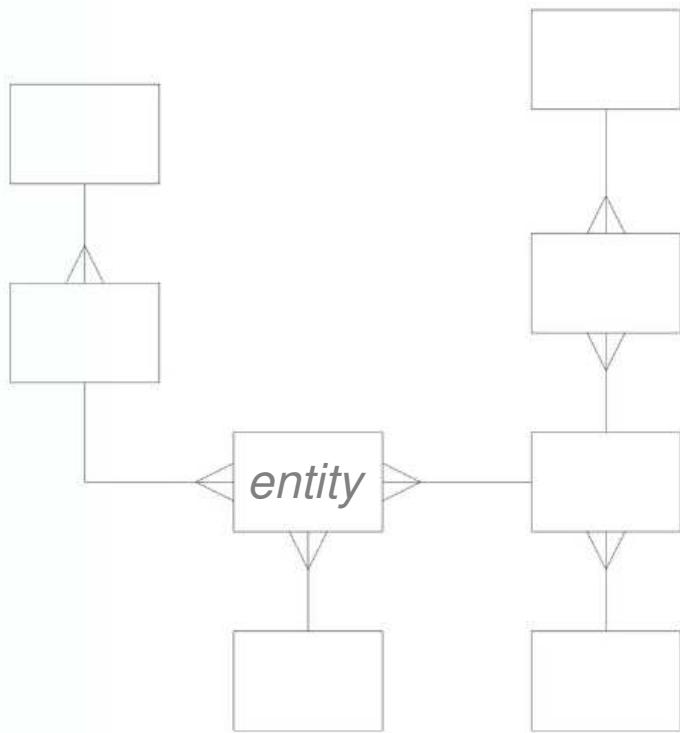
- Transaction processing (OLTP):
 - Fast response time important (< 1 second)
 - Data must be up-to-date, consistent at all times
- Data analysis (OLAP):
 - Queries can consume lots of resources
 - Can saturate CPUs and disk bandwidth
 - Operating on static “snapshot” of data
- OLAP can “crowd out” OLTP transactions
 - Transactions are slow → unhappy users
- Example:
 - Analysis query asks for sum of all sales
 - Acquires lock on sales table for consistency

Порівняння: Data Modeling

- Transaction processing (OLTP):
 - Normalized schema for consistency
 - Complex data models, many tables
 - Limited number of standard queries and updates
- Data analysis (OLAP):
 - Simplicity of data model is important
 - Allow semi-technical users to formulate queries
 - De-normalized schemas are common
 - Fewer joins → improved query performance
 - Fewer tables → schema is easier to understand

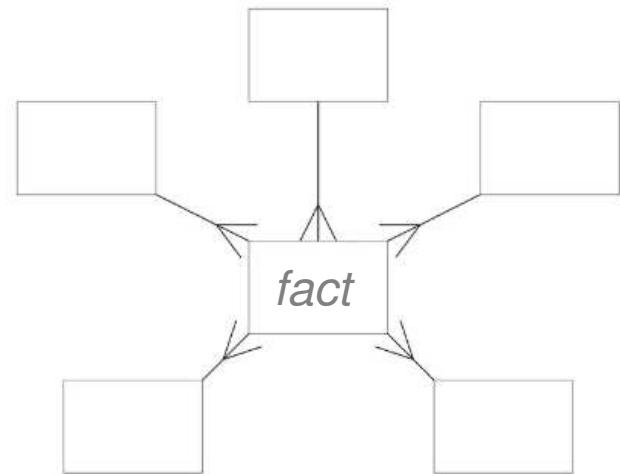
Порівняння: Data Tables

Operational DBMS



ER Diagram

Data Warehouse



Star Schema

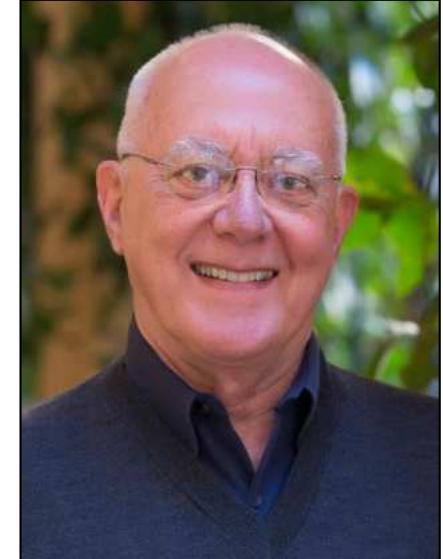
Білл Інмон та Ральф Кімбалл

- **William Inmon (born 1945)** is a computer scientist, recognized as “the father of the data warehousing”, formalized the WH concept
- Received Bachelor degree in mathematics from Yale University in 1967, and his Master degree in computer science from New Mexico State University
- In 1990s introduced the Corporate Information Factory model which proposes an architecture for the integration and management of enterprise data
- In 2007 was named as one of the ten people that most influenced the first 40 years of the IT
- “Building the Data Warehouse” (1992)
- “Architecture for the Next Generation of Data Warehousing” (2008)



Білл Інмон та Ральф Кімбалл

- **Ralph Kimball (born 1944)** is an architect of data warehousing and business intelligence. Ph.D. in 1973 from Stanford University in electrical engineering
- In 1982 developed a graphical programming technique to build a logical flow, allowing a visual style of programming
- Since 1992 has provided data warehouse consulting and education through various companies
- Proposed a “lifecycle methodology” as high-level sequence tasks used to design, develop and deploy DW and BI systems
- “The Data Warehouse Toolkit” (1996)
- “The Data Warehouse Lifecycle Toolkit” (1998)



Правила Едгара Кодда (1993 рік)

- **Multidimensional Conceptual View.** Multidimensional data model is provided that is intuitively analytical and easy to use. It decides how the users perceive business problems.
- **Transparency.** It makes the underlying data repository, computing architecture, and the diverse nature of source data totally transparent to users.
- **Accessibility.** Access should be provided only to the data that is actually needed to perform the specific analysis, presenting a single, coherent and consistent view to the users.
- **Consistent Reporting Performance.** Users should not experience any significant degradation in reporting performance as the number of dimensions or the size increases.
- **Client-Server Architecture.** The system's architecture should be targeted for providing optimum performance, flexibility, adaptability, and interoperability.
- **Generic Dimensionality.** It should be ensured that every data dimension is equivalent in both structure and operational capabilities. Have one logical structure for all dimensions.
- **Dynamic Sparse Matrix Handling.** Adaptation should be of the physical schema to the specific analytical model being created and loaded that optimizes sparse matrix handling.
- **Multi-User Support.** Support should be provided for end users to work concurrently with either the same analytical model or to create different models from the same data.
- **Unrestricted Cross-Dimensional Operations.** System should have abilities to recognize dimensional and automatically perform operations within a dimension or across dimensions.
- **Intuitive Data Manipulation.** Consolidation path reorientation, drill-down, and roll-up to be accomplished intuitively should be enabled and directly via point and click actions.
- **Flexible Reporting.** Business user is provided capabilities to arrange columns, rows, and cells in manner that gives the facility of easy manipulation, analysis and synthesis of information.
- **Unlimited Dimensions and Aggregation Levels.** There should be at least fifteen or twenty data dimensions within a common analytical model.

Властивості FASMI (1995 рік)

- **Fast.** The system is targeted to deliver most responses to users within several seconds, with the simplest analyses taking no more than one second and very few taking less than minute.
- **Analysis.** The system can cope with any business logic and statistical analysis that is relevant for the application and the user, and keep it easy enough for the target user.
- **Shared.** The system implements all the security requirements for confidentiality and, if multiple write access is needed, concurrent update locking at an appropriate level. Not all applications need users to write data back, but for the growing number that do, the system should be able to handle multiple updates in a timely, secure manner.
- **Multidimensional.** The system must provide a multidimensional conceptual view of the data, including full support for hierarchies and multiple hierarchies.
- **Information.** The capacity of various products is measured in terms of how much input data they can handle, not how many gigabytes they take to store it.

<http://dssresources.com/papers/features/pendse04072002.htm>

Визначення сховищ даних

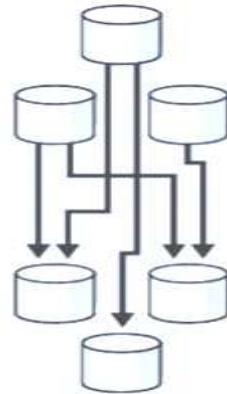
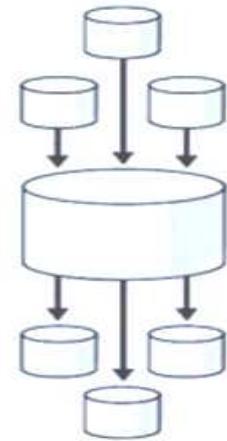
- A subject-oriented, integrated, time variant, non-volatile collection of data in support of management's decision-making process
 - (*Bill Inmon, 1992*)
- A data warehouse is a copy of transaction data specifically structured for querying and reporting.
- An expanded definition for data warehousing includes business intelligence tools, tools to extract, transform and load data into the repository, and tools to manage and retrieve metadata.

Особливості складування даних

- Предметна орієнтація системи
- Інтегрованість даних
- Інваріантність у часі
- Наявність зв'язку із часом
- Висока стабільність даних
- Компромісна надлишковість
- Побудова в термінах бізнес-логіки

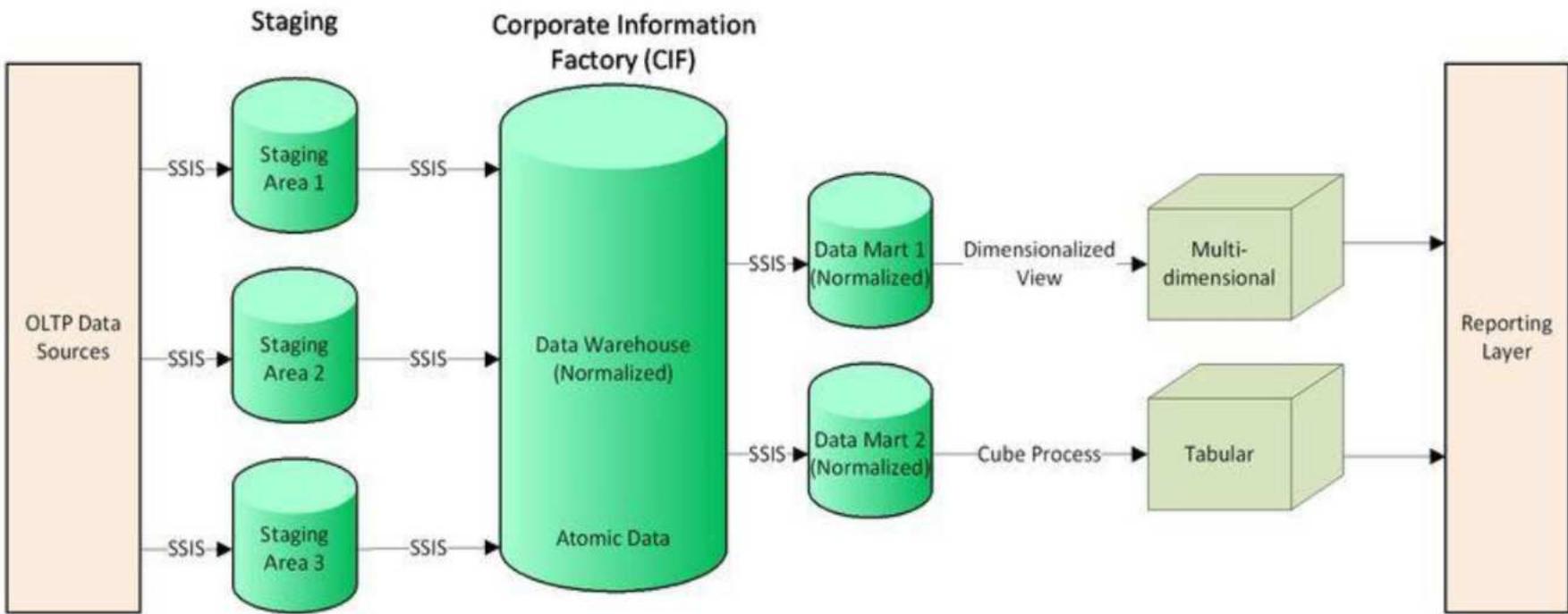
Вітрини або кіоски даних

- Масив спеціалізованої інформації, яка орієнтована на обрану галузь бізнесу або команду
- Операє або локальними даними підрозділу, або отримує данні із глобального сховища
- **Сховище даних:**
 - декілька предметних областей
 - детальна інформація
 - обробляє усі наявні джерела даних
 - аналітична інформація всього підприємства
- **Вітрина даних:**
 - одна предметна область
 - більш стисла інформація
 - ізольовані джерела
 - аналітичні потреби конкретних відділів



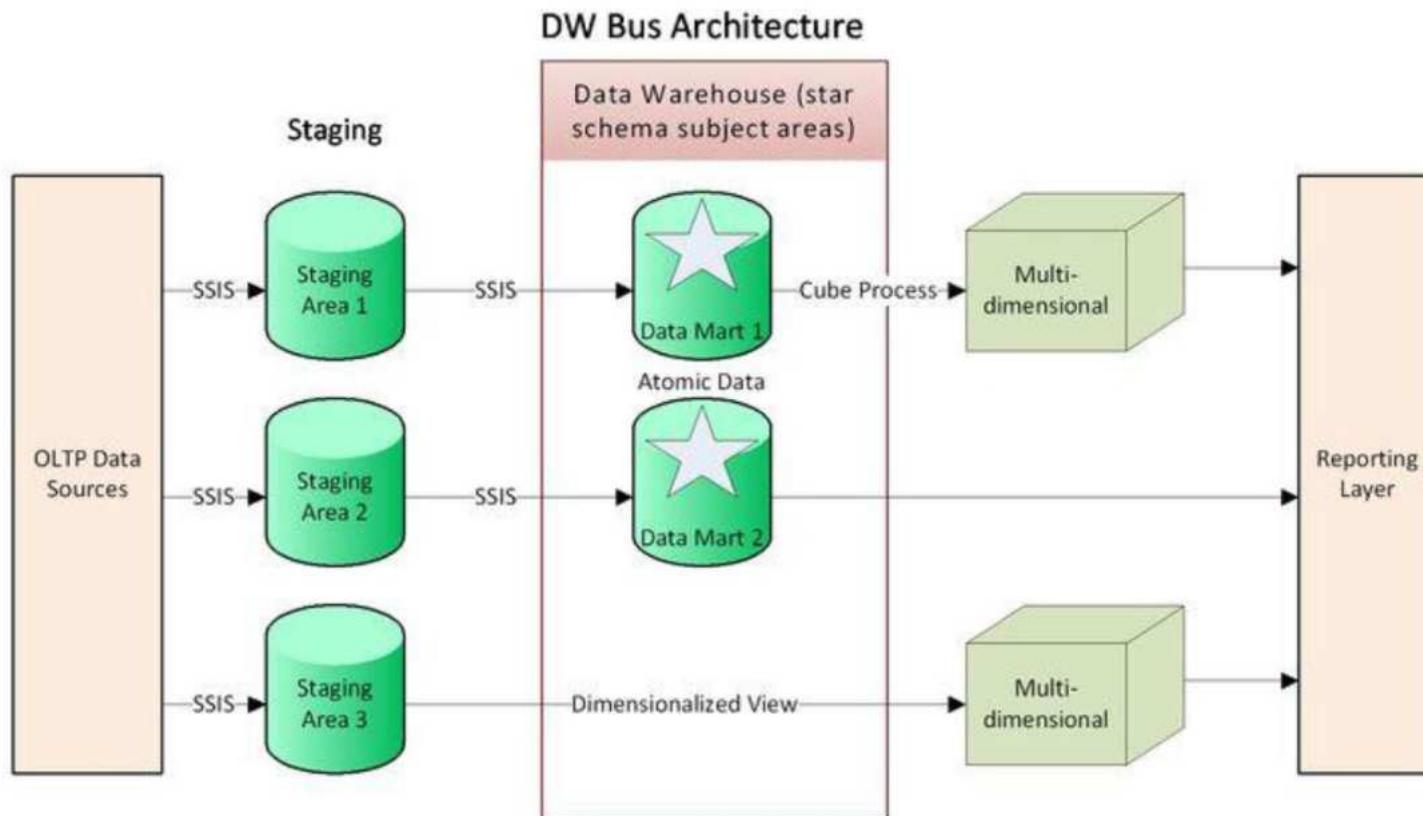
Сховище даних за Інмоном

- Скоординоване отримання інформації із джерел, зберігання багатьох фактів в єдиній узгодженій моделі
- Будується цілісне джерело достовірної, консолідованої інформації для підприємства

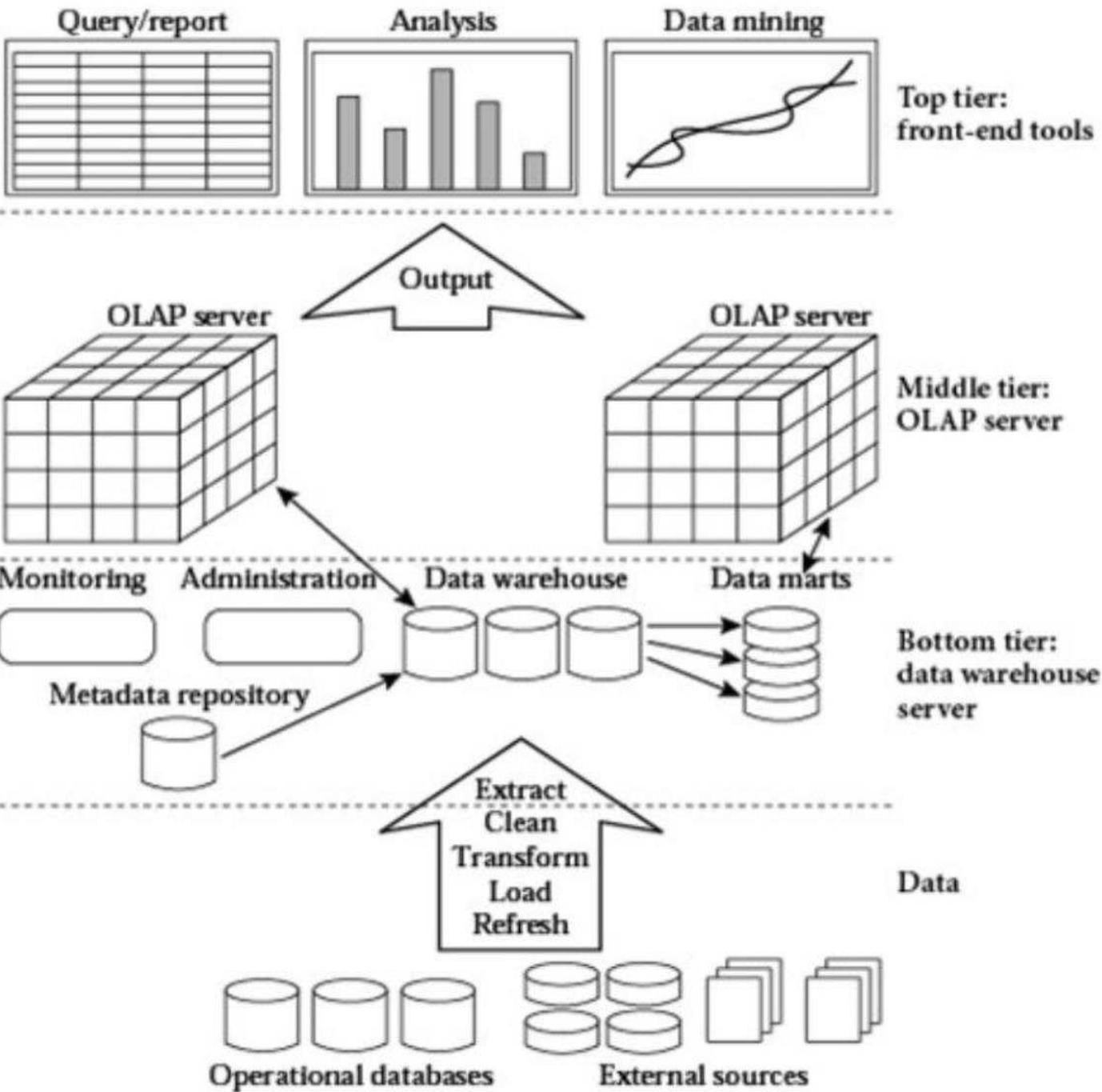


Сховище даних за Кімбалом

- Віртуальне об'єднання автономних вітрин даних, кожна з яких подає один факт
- Операційна бізнес-система виступає джерелом даних, для якого формується аналітична модель



↑
Levels of Data Consolidation



Озера даних – масиви неструктурованої корпоративної інформації в первинній формі

DATA LAKE

vs DATA WAREHOUSE



Raw

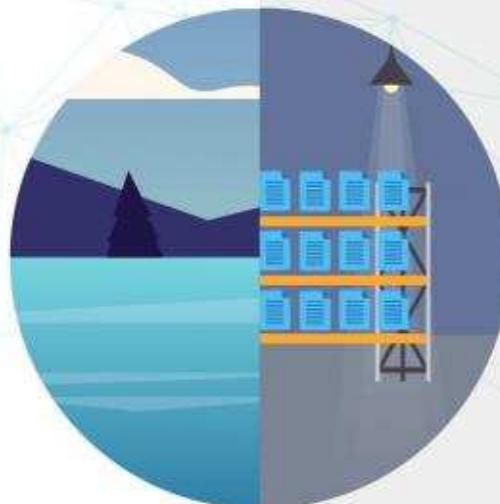
Data Lakes contain unstructured, semi structured and structured data with minimal processing. It can be used to contain unconventional data such as log and sensor data.

Large

Data Lakes contain vast amounts of data in the order of petabytes. Since the data can be in any form or size, large amounts of unstructured data can be stored indefinitely and can be transformed when in use only.

Undefined

Data in data lakes can be used for a wide variety of applications, such as Machine Learning, Streaming analytics, and AI.



Refined

Data Warehouses contain highly structured data that is cleaned, pre-processed and refined. This data is stored for very specific use cases such as BI.

Smaller

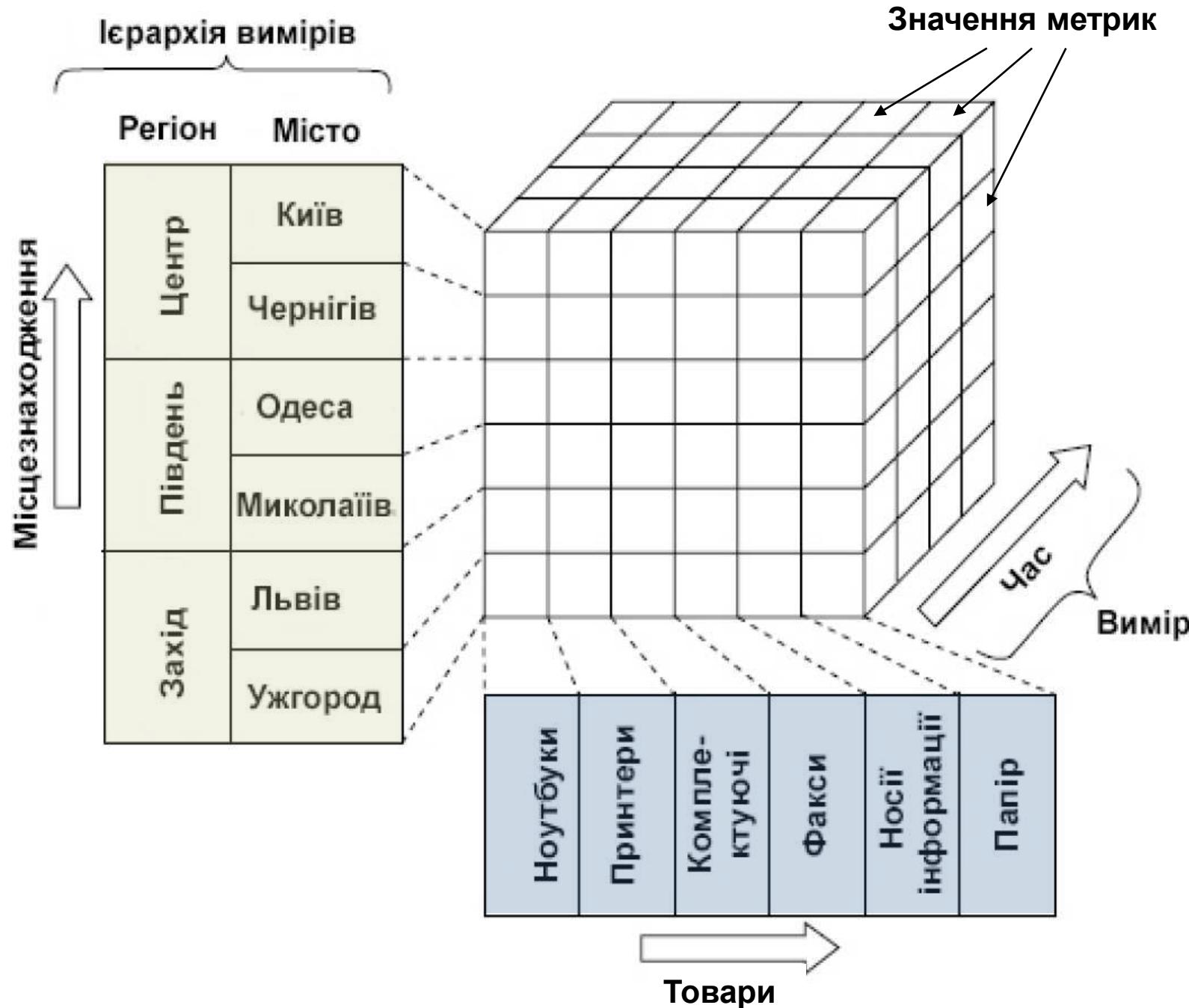
Data Warehouses contain less data in the order of terabytes. In order to maintain data cleanliness and health of the warehouse, Data must be processed before ingestion and periodic purging of data is necessary.

Relational

Data Warehouses contain historic and relational data, such as transaction systems, operations etc.

Багатовимірна модель – засіб накопичення бізнес-інформації для аналізу

- Гіперкуб (**cube**) – абстрактний багатомірний простір, заповнений числовими даними для аналізу
- Факт (**fact**) – набір пов’язаних даних, що чисельно описує здійснену бізнес-операцію
- Атрибут (**attribute**) – опис характеристики об’єкту, який визначається дискретним значенням
- Ось, вимір (**dimension**) – атрибут аналізованого бізнес-процесу, можуть складати ієрархію
- Метрика, міра (**measure**) – числовая характеристика факту, що описує виконані операції з погляду вимірів
- Гранулювання (**granularity**) – рівень деталізації даних, які зберігаються в конкретному кубі



Приклади вимірів та метрик

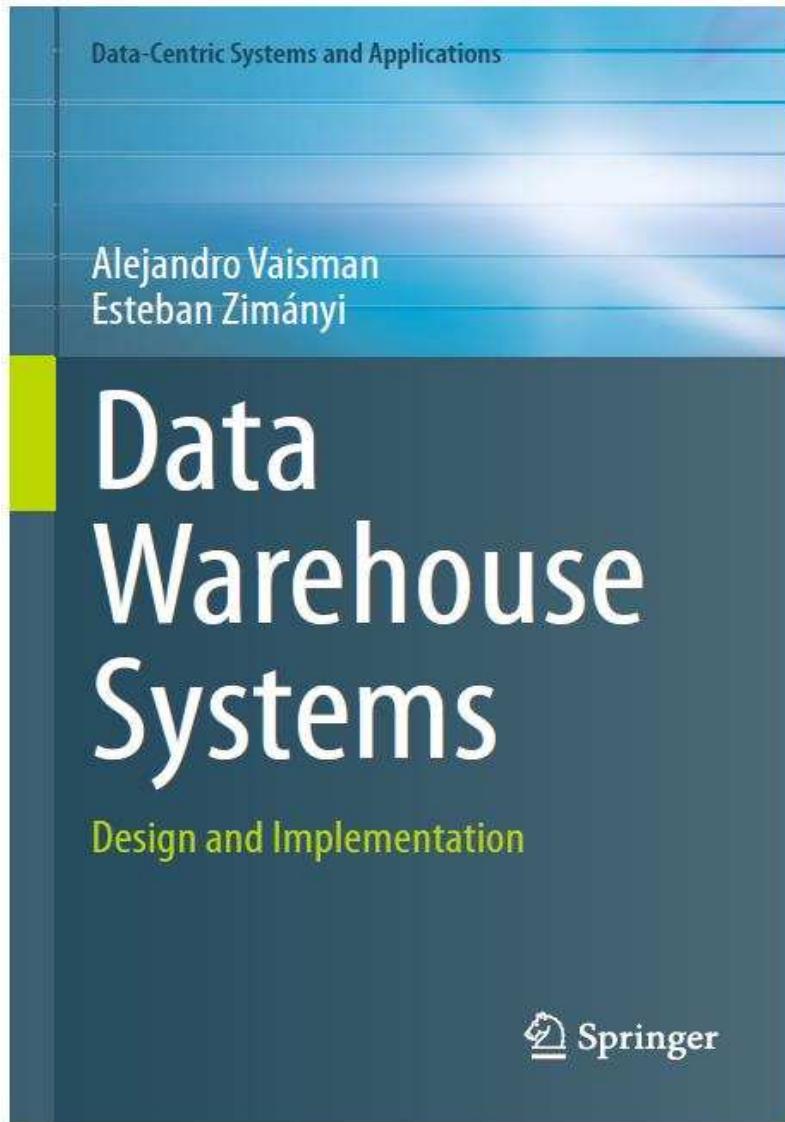
- Ієархія “Місцезнаходження”
 - Місто – Регіон
- Ієархія “Товари”
 - Категорія – Виробник – Країна
- Ієархії “Часу”
 - День – Місяць – Квартал – Рік
 - Дата – День тижня
- Метрики кубу:
 - Ціна продажу
 - Надана знижка
 - Кількість одиниць

<i>Місце:</i>	Київ, Центр
<i>Товар:</i>	Папір, Україна
<i>Час:</i>	сб, 1 бер 2025 р.
<i>Ціна:</i>	500 грн.
<i>Знижка:</i>	20 грн.
<i>Кіль-ть:</i>	3 од.

Інтерпретація даних кубу

- Перетин осей – місце розміщення даних, значення та атрибути вимірів – супровідні дані, решта – метадані сховища
- Хронологія бізнес-події – завжди один із вимірів, який визначається типовою ієрархією за компонентами календаря
- Візуалізація даних – двомірні таблиці, що мають ієрархічні заголовки рядків та (або) стовпців
- Підсумковими значеннями є дані або з вищих рівнів вимірів, або при відкиданні виміру
- З метою пришвидшення зберігаються наперед обчислені агреговані (сумарні, середні) значення для кожного рівня ієрархії

Дякую за увагу!



The Data Warehouse Toolkit

Third Edition

The Definitive Guide
to Dimensional
Modeling

Ralph Kimball
Margy Ross



СХОВИЩА ДАНИХ: Лекція №10

НУ “Львівська Політехніка”, кафедра ПЗ

Програмне забезпечення для
реалізації сховищ даних

Процеси реалізації сховища даних

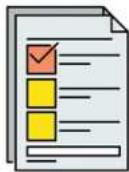
- Ідентифікація бізнес-інформації, яка має зберігатися (сущності, залежності, повнота, гнучкість структури)
- Аналіз джерел даних із точки зору якості, узгодженості, часової природи (варіативність, періодичність, грануляція)
- Розрахунок обчислювальних потужностей, прогнозованих обсягів даних, частота та складність запитів
- Виявлення ролей користувачів та політик обмеження доступу
- Моделювання сценаріїв отримання аналітичної інформації
- Затвердження словника для взаємодії різних груп користувачів
- Обрання технологій, програмного забезпечення та додатків
- Встановлення процедур оновлення даних, які відповідають часовим і життєвим циклам бізнес-інформації
- Розробка інтерфейсів моніторингу системи та під'єднаних процесів отримання, завантаження даних, швидкості обробці запитів
- Опис підходів до оформлення технічної документації
- Техніко-економічне обґрунтування розробки та супроводу

Планування
Проектування

→ Збір вимог
→ Побудова

→ Аналіз
→ Впровадження

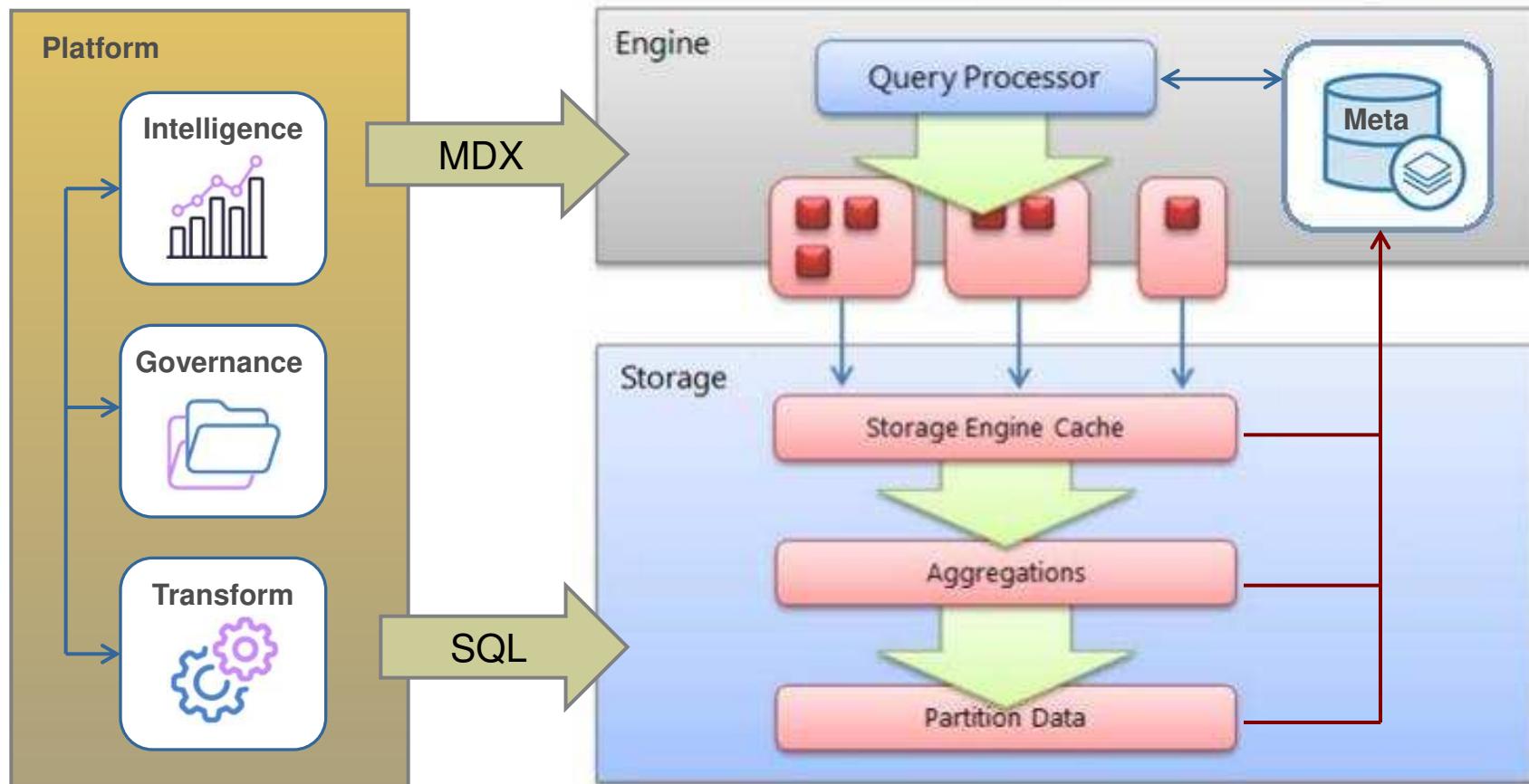
Склад системи бізнес-аналітики



Business Analytics Requirements
Data Governance Procedures
Security Policy Specifications



Efficiency Requirements
Technical Infrastructure
Reliability and Failures Monitors



Призначення метаданих сховища

- Опис відповідності джерел даних моделям сховища
- Інтерпретація в термінах бізнес-користувачів
- Забезпечення відкритості даних для інших систем
- Історія внесення даних у сховище, повнота, верифікація
- Інформаційна безпека, обмеження доступу, керування
- **Рівні стандарту CWM v1.1 від 2003 року:**

	Warehouse Process			Warehouse Operation		
	Transformation		OLAP	Data Mining	Information Visualization	Business Nomenclature
Management						
Analysis						
Resources	Object Model	Relational	Record	Multidimensional		XML
Foundation	Business Information	Data Types	Expression	Keys and Indexes	Type Mapping	Software Deployment

Стандарт формалізації метаданих

- Стандарт **CWM** визначає модельно-орієнтований підхід до обміну метаданими, записаними у форматі **UML**, відповідно до синтаксичних і семантичних характеристик оброблюваної бізнес-інформації
- Використовує **XMI** серіалізацію для забезпечення інтер-operабельності між різними системами



Шари метаданих

- **Foundation** (основа) – класи та асоціації, які формують ядро об'єктної моделі, описують поведінку об'єктів, визначають відповідні виклики, задають формати і типи даних
- **Resources** (ресурси) – пакети для опису зв'язаних інформаційних джерел і цільових баз даних
- **Analysis** (аналіз) – моделювання процесів і служб інформаційного аналізу, включаючи візуалізацію і поширення даних, вилучення і видобування знань, багатомірний аналіз та ін.
- **Management** (управління) – засоби функціонування сховища даних, моделювання процедур управління, встановлення регламенту їх виконання, деталізація процесів контролю і протоколювання етапів обробки даних

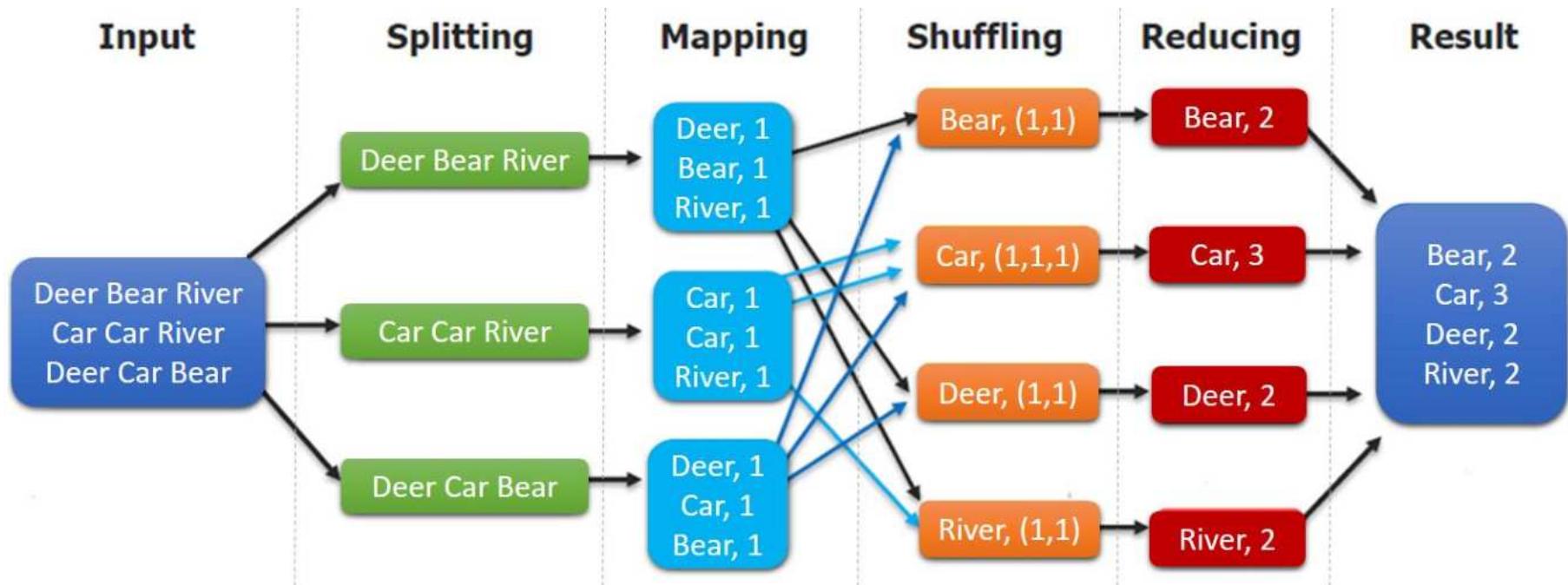
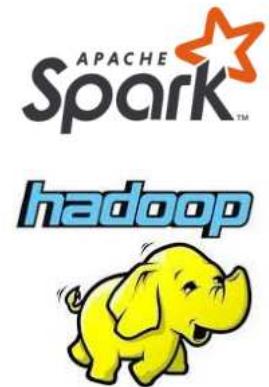
Зберігання багатовимірних даних

- **Таблиця фактів має значну кількість рядків**
- Не всі колонки потрібні для кожного запиту
- Статичність даних, природні партіції за часом
- Групування на довільних наборах ключів вимірів
- Відсутнє сортування по даним, що зберігаються

- **Виміри мають нереляційну ієрархічну модель**
- Обмежена кількість дискретних значень
- Множина ключів для операції зрізу кубу зазвичай має невеликий розмір
- Багатократне дублювання даних вимірів
- Типові вибірки не виконують складних з'єднань

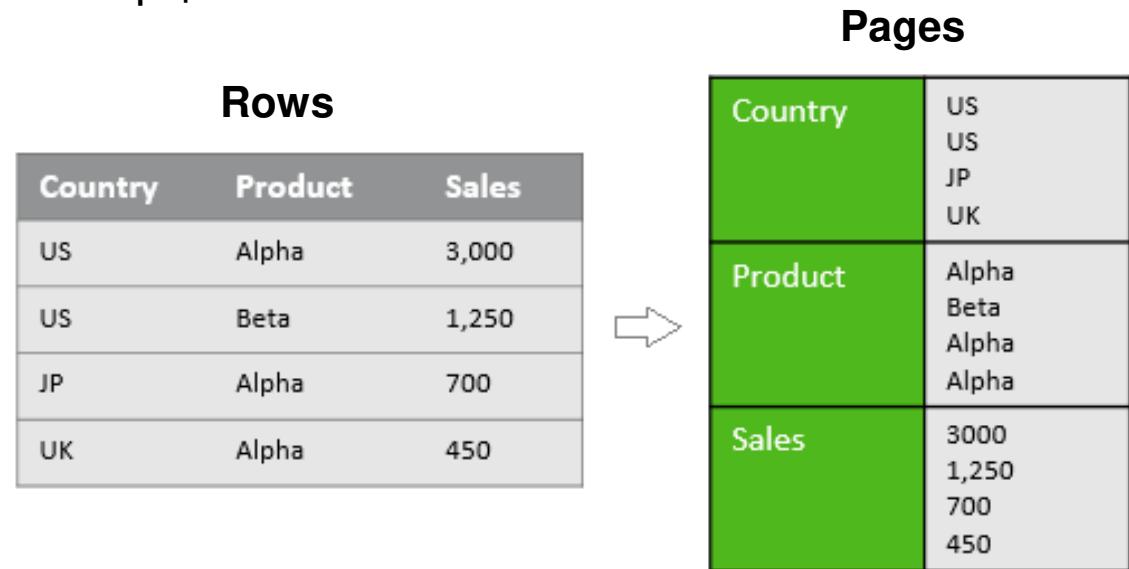
Концепція кластеру Map-Reduce

- Розподілена система зберігання, асинхронна реплікація
- Паралельне виконання операцій на вузлах кластеру
- Розбиття запиту на потоки, призначення задач, контроль
- Трансформація вхідних даних в асоціативний масив, де проміжні значення пов'язані із одним і тим же ключем
- Збирання даних від вузлів в цілісний кінцевий результат



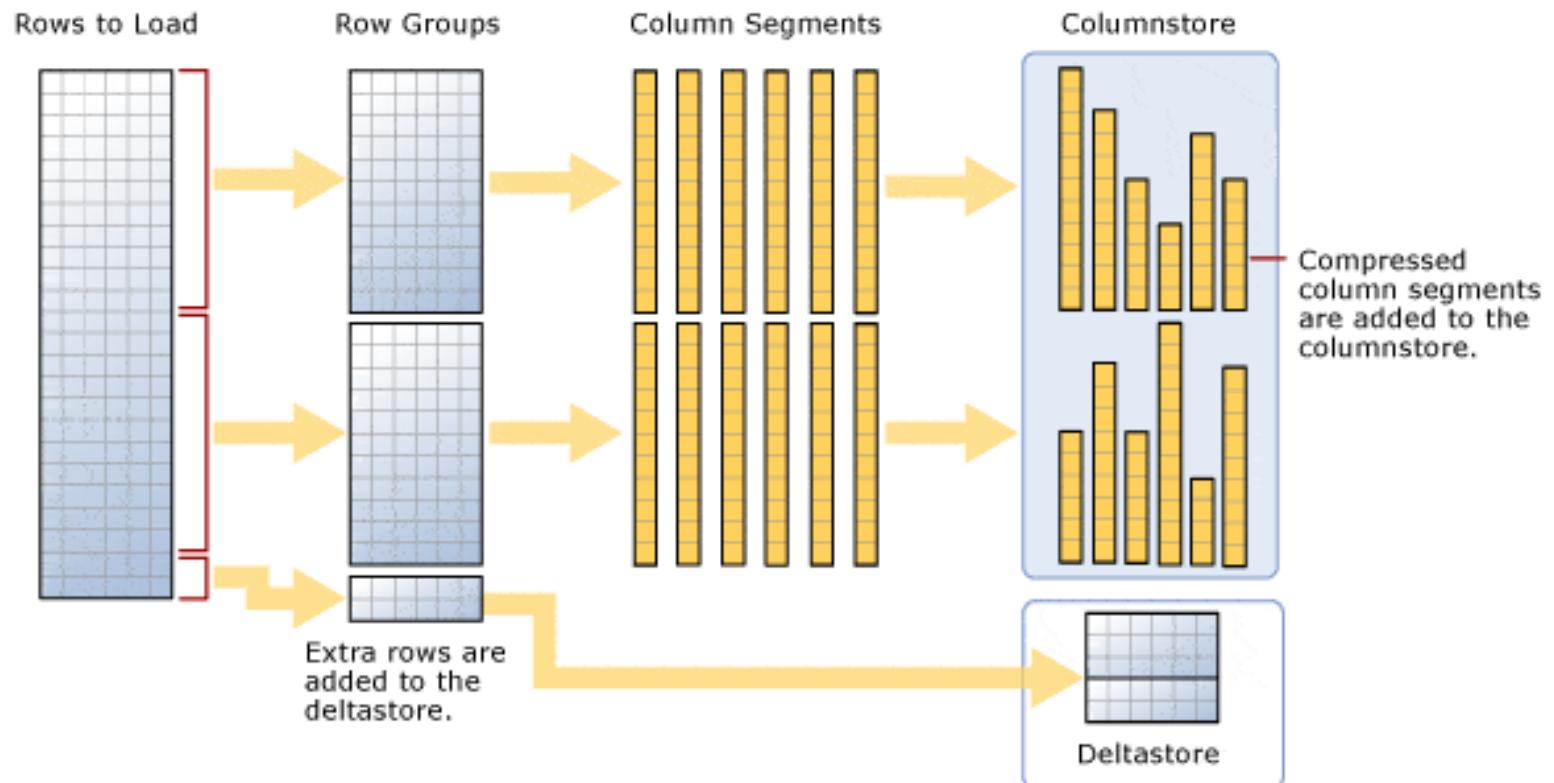
Бази даних типу Column-Store

- Дані кожної колонки зберігаються послідовно у відповідних сторінках як масив заданого типу
- Розбиваються по блоках на фіксовану кількість рядків, що адресуються за порядковим номером
- При формуванні результатів вибірки читаються сторінки тих колонок, які потрібні для агрегації і відповідають умовам фільтрації
- Однорідна природа даних у блоці дозволяє ефективно використовувати компресію
- Паралельне виконання операцій пошуку завдяки тому, що дані вже розбити на порції



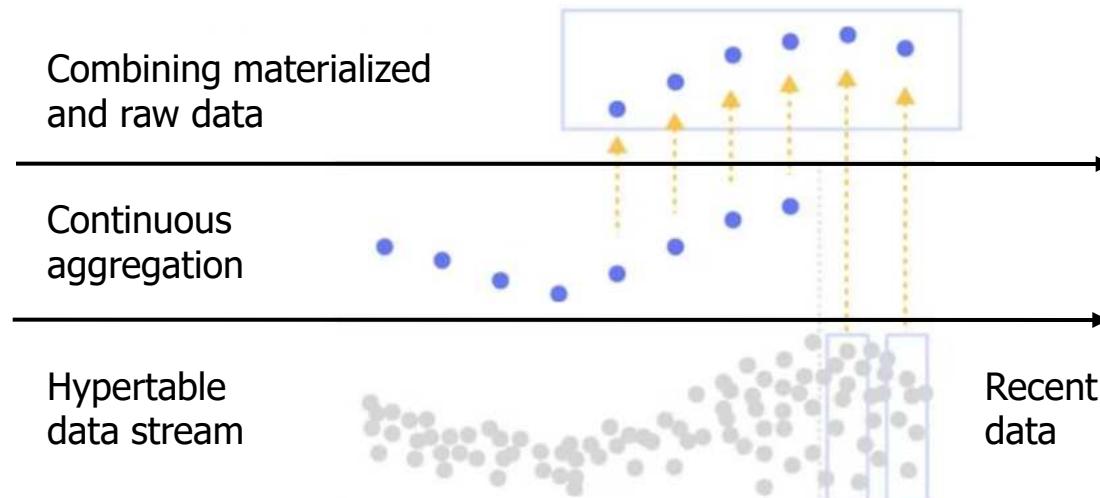
Кластеризовані індекси Column-Store

- Компактна організація даних колонок в рядкових СУБД
- Розбиття на сегменти для паралельної обробки
- Додаткове партіціювання таблиць зменшує об'єми, необхідні для зчитування в межах однієї вибірки
- Зміни накопичуються окремо для фонової перебудови



Aggregate-індекси та представлення

- Зберігають обчислені значення деяких функцій (SUM, MIN, MAX) при групуванні на обраних колонках
- Поступово накопичують зміни даних, що надходять до таблиці після побудови індексу-агрегату
- Поточні зміни враховуються при обчисленні значення



- Аналогом є матеріалізоване представлення (MATERIALIZED VIEW) із звичайним індексом, задане по запиту з агрегатними функціями
- На відміну від індексів-агрегатів його використання потребує підсистеми перенаправлення запитів
- Підтримка інкрементального оновлення даних представлень та автоматичне використання обчислених даних у запитах GROUP BY

Бітмар-індекси для ключів фактів

- Кожне значення для поля-ключа має власну бітову карту по номерах рядків в блоках
- Розрідженість даних (багато нулів в карті) надає швидку компресію
- Оператори NOT, AND, OR для будь-якої комбінації колонок або значень
- Агрегація метрик за індексом не потребує складних операцій
- Зберігається на кінцевих вузлах бінарного дерева
- Потребує оновлення всіх вузлів блоку при внесенні

≥ Green

≥ Red

≥ Blue

ID, Color	Red	Green	Blue	Black
1, Red	1	0	0	0
2, Blue	0	0	1	0
3, Black	0	0	0	1
4, Green	0	1	0	0
5, Red	1	0	0	0
6, Black	0	0	0	1
7, Black	0	0	0	1
8, Red	1	0	0	0
9, Blue	0	0	1	0
10, Green	0	1	0	0
11, Red	1	0	0	0
12, Blue	0	0	1	0
13, Black	0	0	0	1

Поєднання індексів в операції зрізу

- Створення інверсного індексу для заданої колонки таблиці, з'єднаної за зовнішнім ключем:
 - CREATE BITMAP INDEX ON Sales([Products.Category](#)) FROM Sales, Products WHERE Sales.pID = Products.pID
- Поєднання умов відбувається побітовою операцією:
 - SELECT * FROM Sales JOIN Products JOIN Clients WHERE Category = '[Food](#)' AND City = '[Eindhoven](#)'
 - Bitmap(**0011**) & Bitmap(**0110**) → Bitmap(**0010**) ⇔ Row(**№3**)

ORACLE
teradata.

IBM DB2

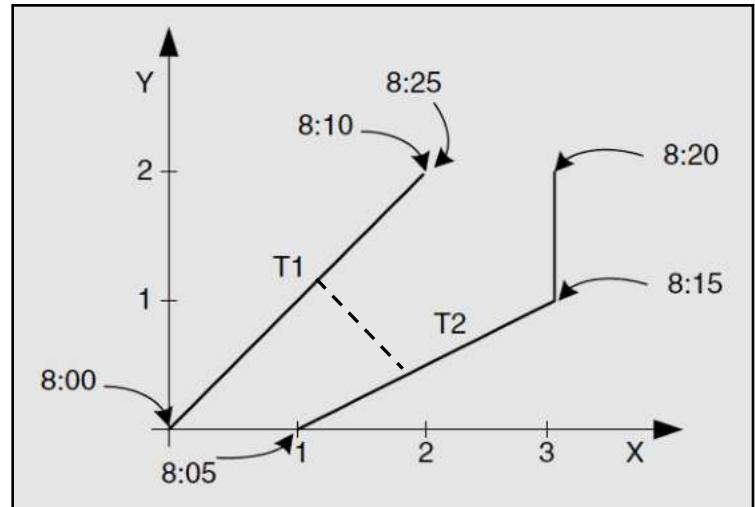
Sales			Products				Clients		
	Date	pID	Client	pID	pName	Category	Price	Customer	City
№1	10/5/12	1	Jack	1	Jacket	Non-food	10	Jack	Brussels
№2	10/5/12	1	Pete	2	Bread	Food	2.1	Mary	Brussels
№3	13/5/12	3	John	3	Beer	Food	1.5	John	Eindhoven
№4	14/5/12	2	Mary	4	Paper	Non-food	1.2	Pete	Eindhoven

Category	Bitmap
Non-food	1100
Food	0011

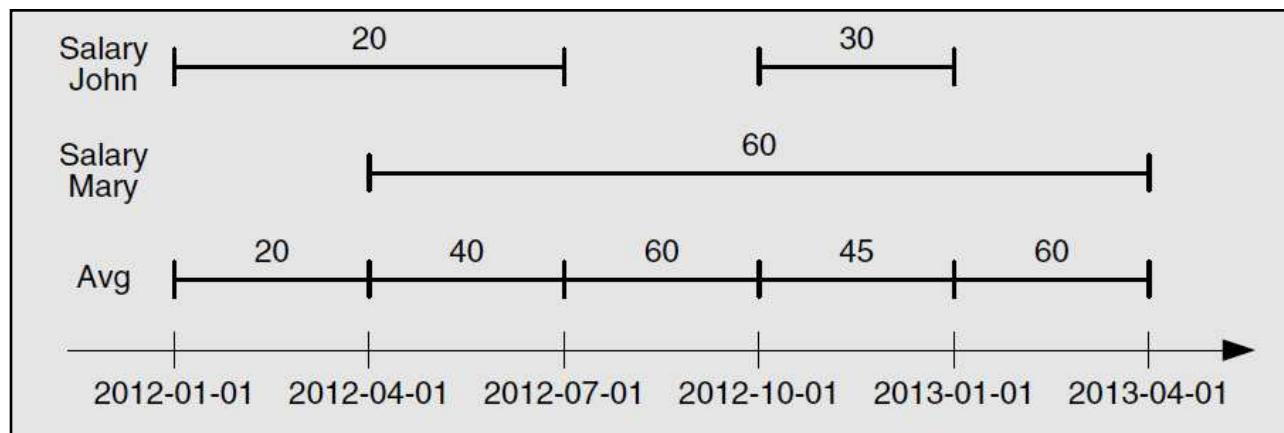
City	Bitmap
Brussels	1001
Eindhoven	0110

Аналітика довготривалих процесів

- Встановлення епохи актуальності даних
- Обробка повільно змінних вимірів
- Збереження метрик для відрізку часу
- Непереривна функціональна залежність між дискретними значеннями вимірів
 - Vary both on time and space
- Спеціальна алгебра для часових запитів
- Статистичні функції агрегації даних
 - Rate of change (Derivative, Speed, Turn)
 - Temporal aggregation (Integral, Duration, Length, Average, Variance, StdDev)



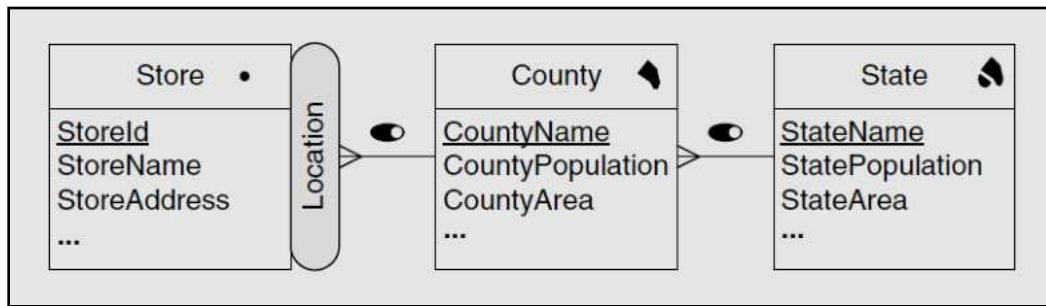
Route Distance (T1, T2)



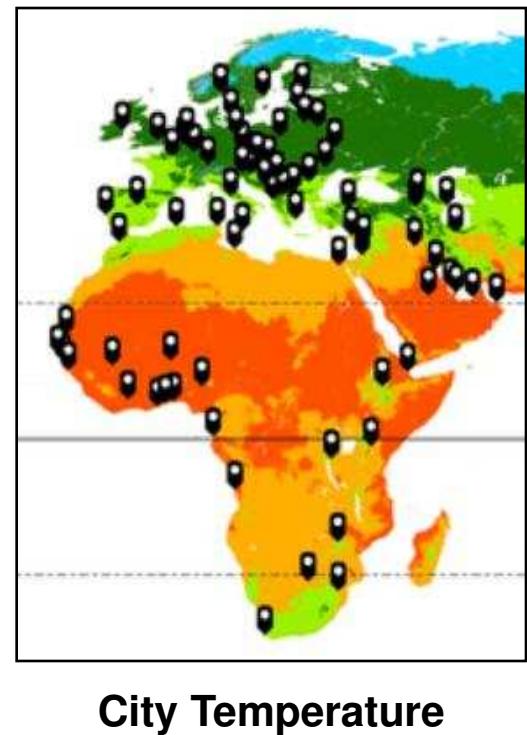
**Temporal Average
(John, Mary)**

Аналітика просторових даних

- Аналіз за формою сущності (координатами, контуром), а не за текстовим атрибутом (адресою, містом)
- Ієрархії вимірів мають опис просторового регіону
- Метрики для геометричних об'єктів реального світу
- Відповідність точок об'єкту певному числовому значенню



- Спеціальні оператори та агрегатні функції:
 - Topological relationships (Overlaps, Disjoint, Equals, Contains, Covers, Touches, Crosses)
 - Numeric operations (Components, Length, Area, Perimeter, Distance, Direction)
 - Operations on field (Integral, Area, Surface, Average, Variance, StdDev)



Граф-орієнтована модель сховища

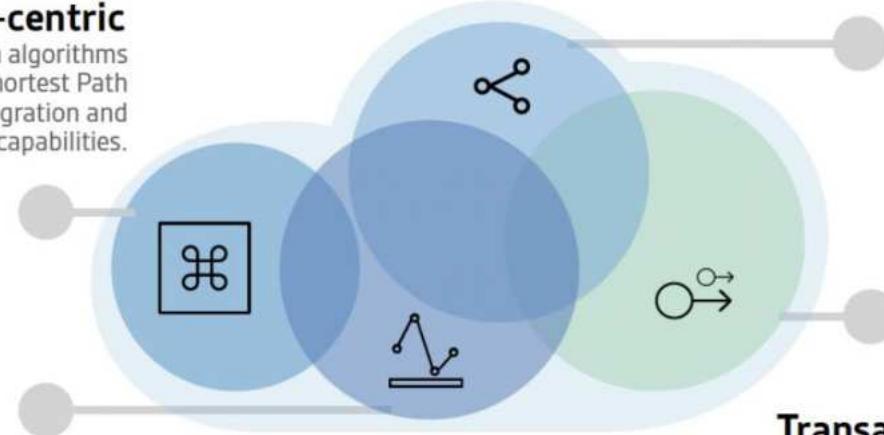
- Подання бізнес-інформації із гнучкою структурою зв'язків
- Зрізи із різних точок зору на різних рівнях грануляції
- Фактами є окремі графи-знімки із актуальними відношеннями
- Агрегація поєднує знімки по ребрах, вершинах
- Виміри: Informational, Topological (локальний стан зв'язків)
- Метрики: Informational, Structural

Semantic

Focused on ontologies and semantic reasoning with a nod toward Tim Berners-Lee and internet data sharing. Good at data harmonization, but may lack analytical powers granted by property graphs.

Algorithm-centric

Great at running graph algorithms like PageRank and Shortest Path but may lack data integration and harmonization capabilities.



Analytical

Offer a good balance for semantic, algorithmic and data warehouse style analytics. Deep analysis on large data sets.

Transactional

Talented at storing away data and running short-running queries like lookups, but may fall short on deep analytics that traverse the graph.

Apache Kylin – Leading open source engine for Big Data



- **Unified Data Analytics Platform**
 - Data analytics from different platform including Hadoop, Cloud, RDBMS, while providing a unified interface for downstream applications
- **Ready to Business Intelligence**
 - Support connecting to different BI tools like Tableau/Power BI/Excel using MDX endpoint connector
- **Brand New Front-End**
 - Define table relationships, dimensions and measures in a single canvas during a modeling process

- **Fault-Tolerant Warehouse System**
 - Enables analytics at a massive scale and facilitates reading, writing, and managing petabytes of data residing in distributed storage using SQL
- **Built-In Metadata Infrastructure**
 - Metastore provides two important features of a data warehouse: data abstraction and data discovery
 - Data discovery enables users to discover and explore relevant and specific data in the warehouse

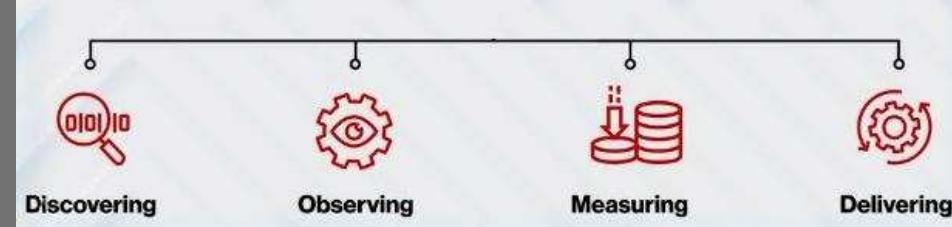


An American nonprofit corporation to support a number of open-source software projects. It represents world-wide decentralized community of the software developers

Pentaho – Platform for complete data management

Pentaho

HITACHI
Inspire the Next



Data Integration

- Ingest, blend, cleanse and prepare diverse data from any source
- Codeless orchestration that blends data sets into a place of truth
- Intuitive, drag-and-drop designer to simplify the creation of data pipelines
- Coordinate, combine transformations with notifications and alerts
- Automatic publishing of metadata models to drive faster results

Data Catalog

- Find, analyze and tag structured and unstructured data
- Contextualize with business glossary and governance policies
- Identifying data anomalies, errors, and outliers
- Use analytics models to predict data quality issues
- Organize and archive data by lifecycle stage

Business Analytics

- Rich library of interactive visualizations with attribute highlighting
- Ability to drill down into supporting reports for detailed data
- Deploy rich dashboards through a drag-and-drop interface
- Operational and parameterized reports, interactive self-service reporting



**+ Open Source
Community Edition**

Microsoft – Data warehousing and analytics software

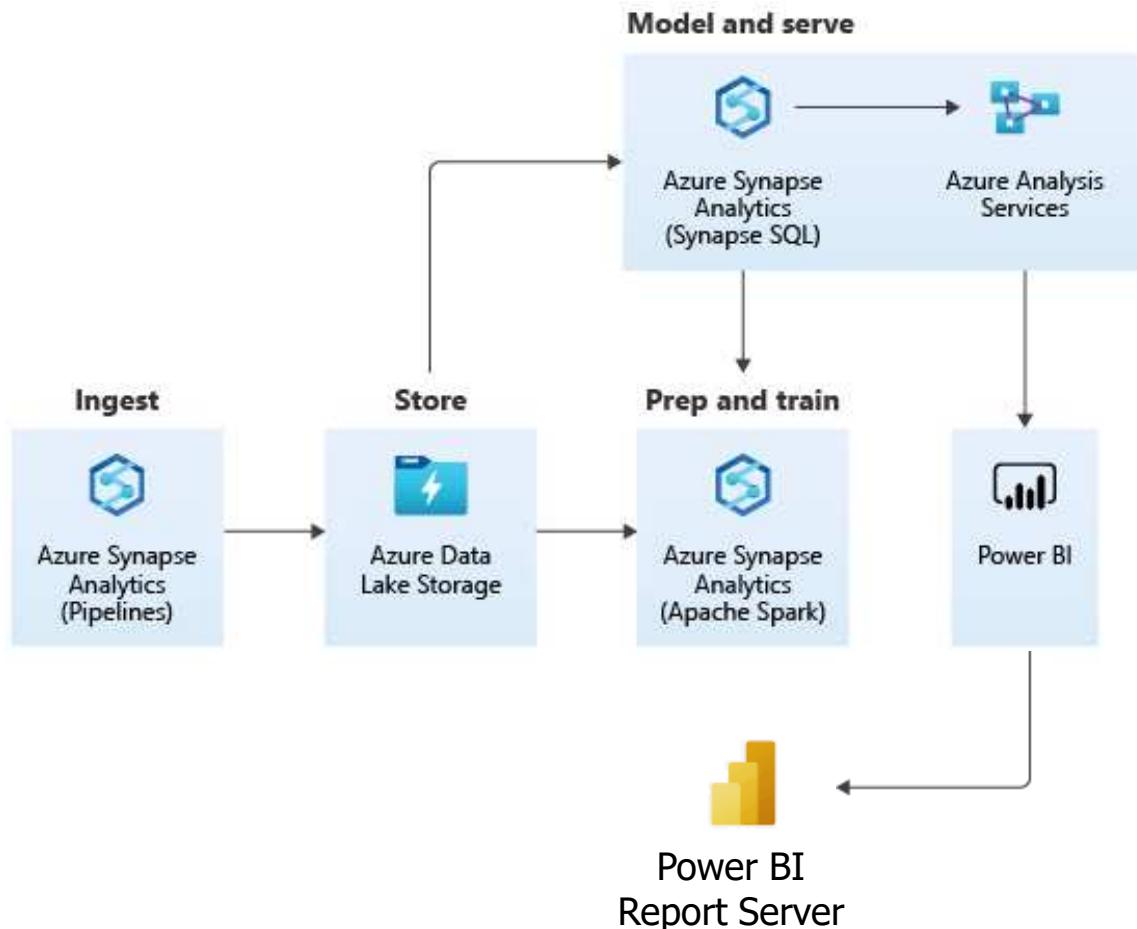


SQL Server Data Tools

- Analysis Services
- Integration Services
- Reporting Services

SQL Package CLI

- CI/CD pipelines
- Automate build and deployment of database projects



Microsoft Power BI

- Desktop
- Report Builder

Qlik – Integrate, transform, analyze and act on data



Streaming Data with Change Data Capture

Deliver volumes of real-time, analytics and AI-ready data.



Data Warehouse Automation

Accelerate and simplify the data warehouse lifecycle.



Data Lake / Data Lakehouse Creation

Enable the agile data lake.



Data Quality and Governance

Find, manage, and fix your data quickly and securely.



Augmented Analytics

Make insights accessible to everyone with automated insight generation, predictive analytics, and generative AI.



Embedded Analytics

Your business users can make better decisions faster with real-time insights and information embedded in apps they already use.



Visualizations and Dashboards

No matter the technical skill, you can quickly search and explore contextually and interactively across all datasets, in any direction.



Reporting

Automated, no-code report templates and options work with all major platforms and apps.

Промислові вендори із власними СУБД

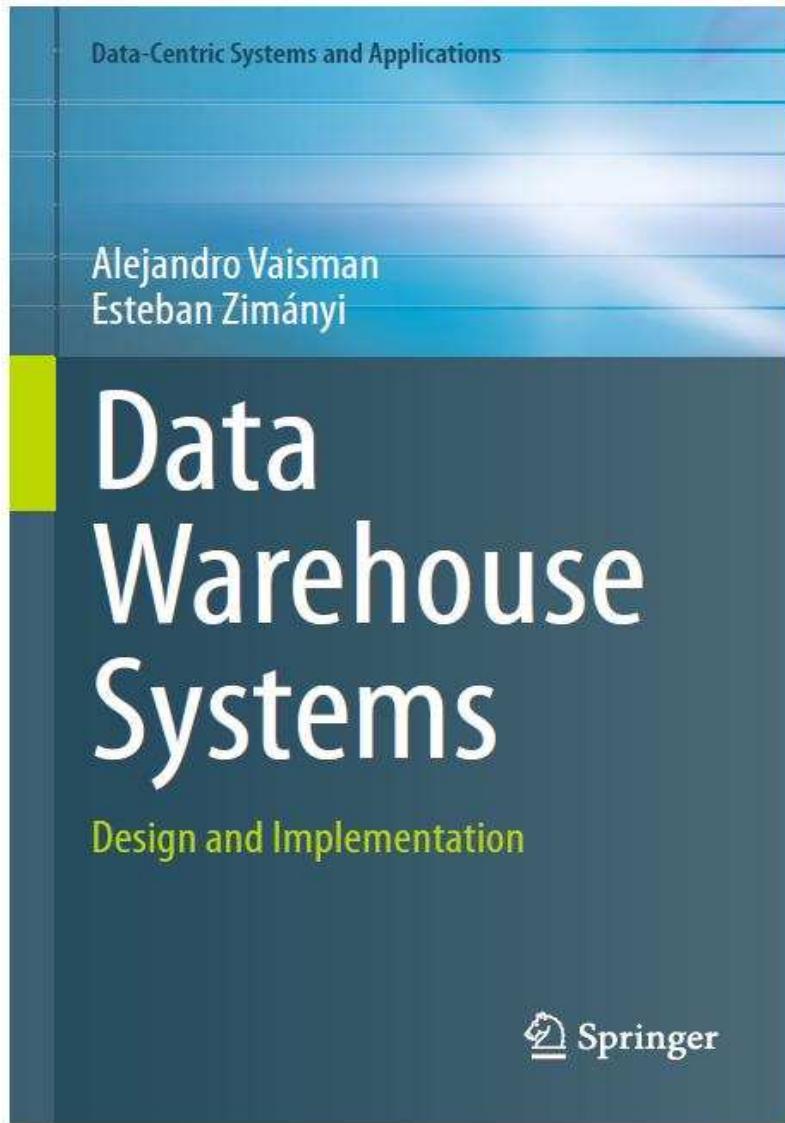
- **Oracle OLAP Technology**
 - Analytic workspace manager software
 - Cube-aware OLAP DML statements
 - Measure folders to simplify access for end users
- **Microsoft Server Analysis Services**
 - Different OLAP providers and data sources
 - Querying, BI reporting and data integration services
 - Interfaces to manipulate multidimensional databases
 - Attribute relationships, star and snowflake schemas
- **SAP Business Objects Analysis**
 - Different OLAP data source objects
 - Totals, parents and aggregations
 - Analysis workspaces and sheets
- **IBM Cognos Analytics**
 - IBM Analysis Studio, Cognos Cube Designer
 - Cognos Dynamic Cubes keeps dimensional metadata
 - Powerplay and Transformer for customized data extraction
- **SAS Intelligence Platform**
 - SAS OLAP Server with metadata that defines cubes, cube data, access permissions and load-balancing details
 - SAS Cube Studio is a viewer generates MDX-bases queries
 - SAS Information Map Studio



Платформи аналітичної обробки даних

- **Cloud Platforms** with enhanced BI facilities
 - Qlik Sense, Tableau, SnowFlake, MicroStrategy, Sisense, Looker, Domo, Zoho, Wunderdata, ...
- **Google BigQuery** is a cloud data warehouse
 - Table data storage with snapshots and materialized views
 - External tables where the data resides in a cloud
 - Ad hoc workflows analysis, analyze and visualize geospatial data, execute machine learning models
 - Business intelligence is a fast analysis service to build rich, interactive dashboards
- **Apache Drill** is an open-source SQL query engine
 - Query any non-relational datastore
 - Treat data like a table even when it's not
 - No centralized metadata requirements
 - Columnar execution engine
 - Data-driven compilation and recompilation

Дякую за увагу!



The Data Warehouse Toolkit

Third Edition

The Definitive Guide
to Dimensional
Modeling

Ralph Kimball
Margy Ross

