# Сховище даних

1. Концепція СД.
2. Організація СД. Типи даних.
3. Очистка даних.

# СХОВИЩЕ ДАНИХ (СД)
## Data Warehouse

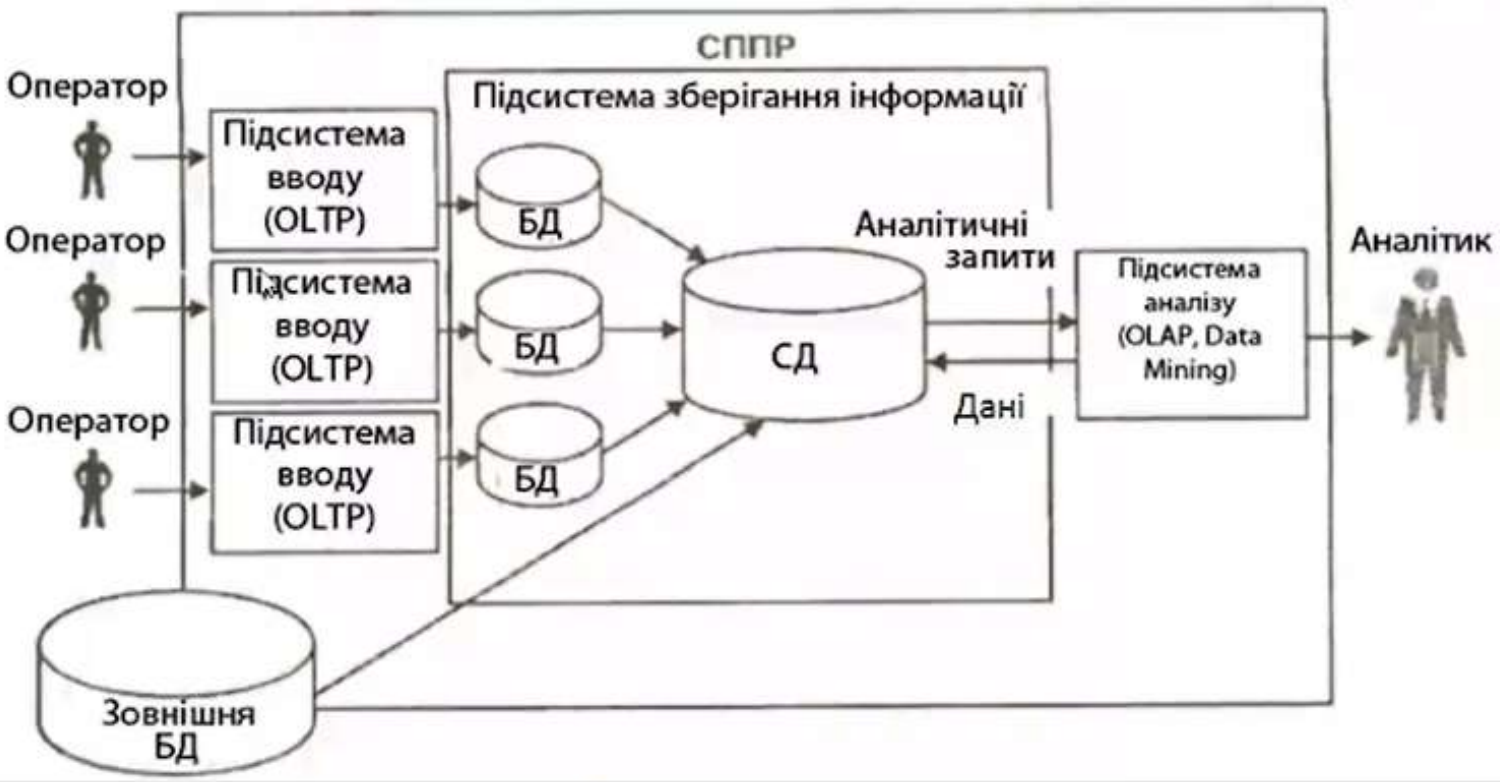
**В 1992 році William H. Inmon детально описав концепцію СД у своїй монографії «Побудова сховищ даних» («Building the Data Warehouse»), де дав наступне означення.**

***Сховище даних*** **– предметно-орієнтований, інтегрований, незмінний** ***набір даних*****, такий що підтримує хронологію і організований для цілей підтримки прийняття рішень.**

# Основні властивості Сховища Даних

- Предметна орієнтація (дані отримані з однієї БД можуть подавати хибні результати аналізу)
- Інтеграція (всі дані повинні бути зведені до єдиного формату)
- Підтримка хронології (всі дані повинні зберігатись в хронологічному порядку, це впливає на аналіз)
- Незмінність (з БД дані видаляють і оновлюють, з СД – ні)

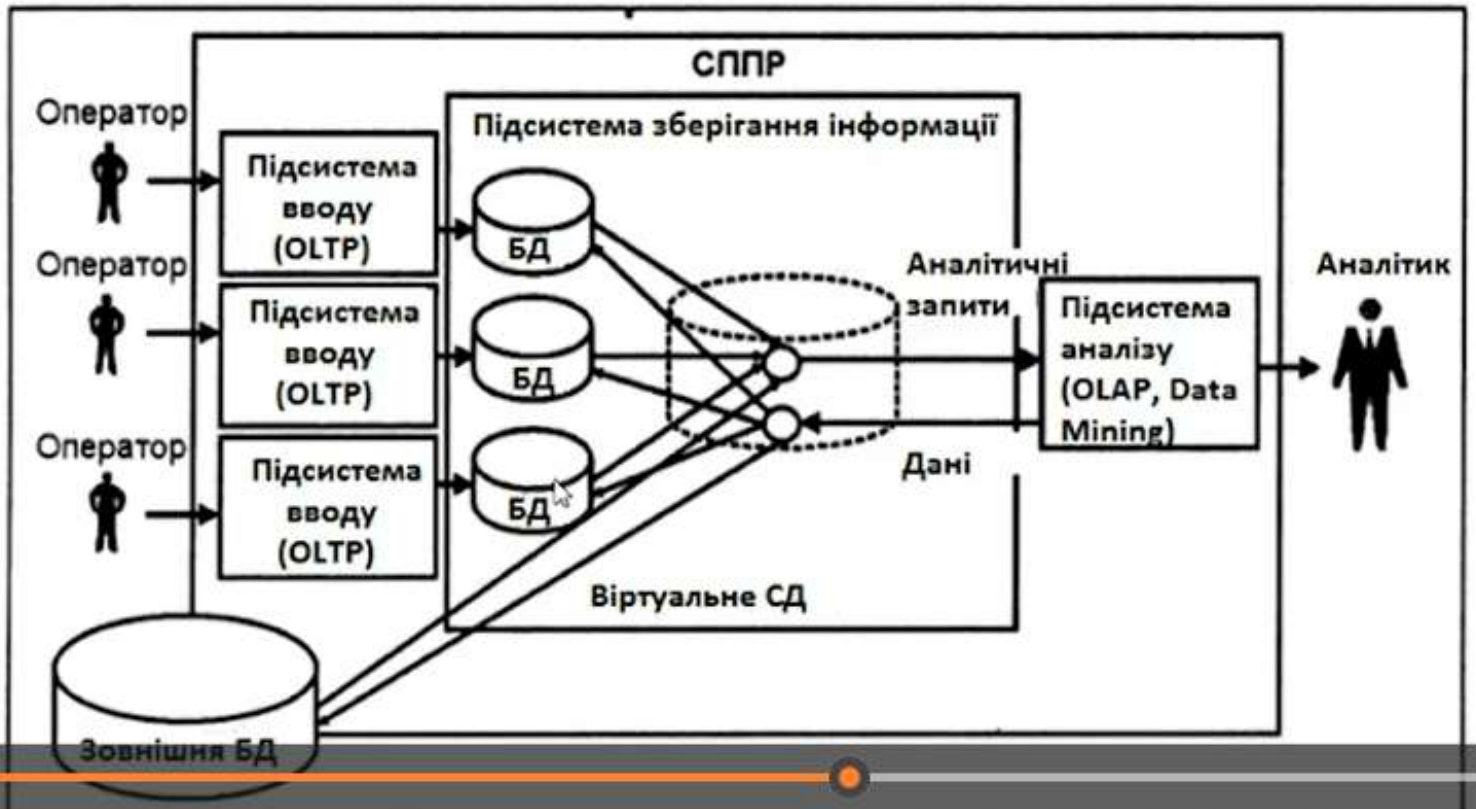
# Структура СППР з фізичним (класичним) СД




ОДД – оперативне джерело даних

OLTP – Online transaction processing; OLAP – Online analytical processing

# Структура СППР з віртуальним СД

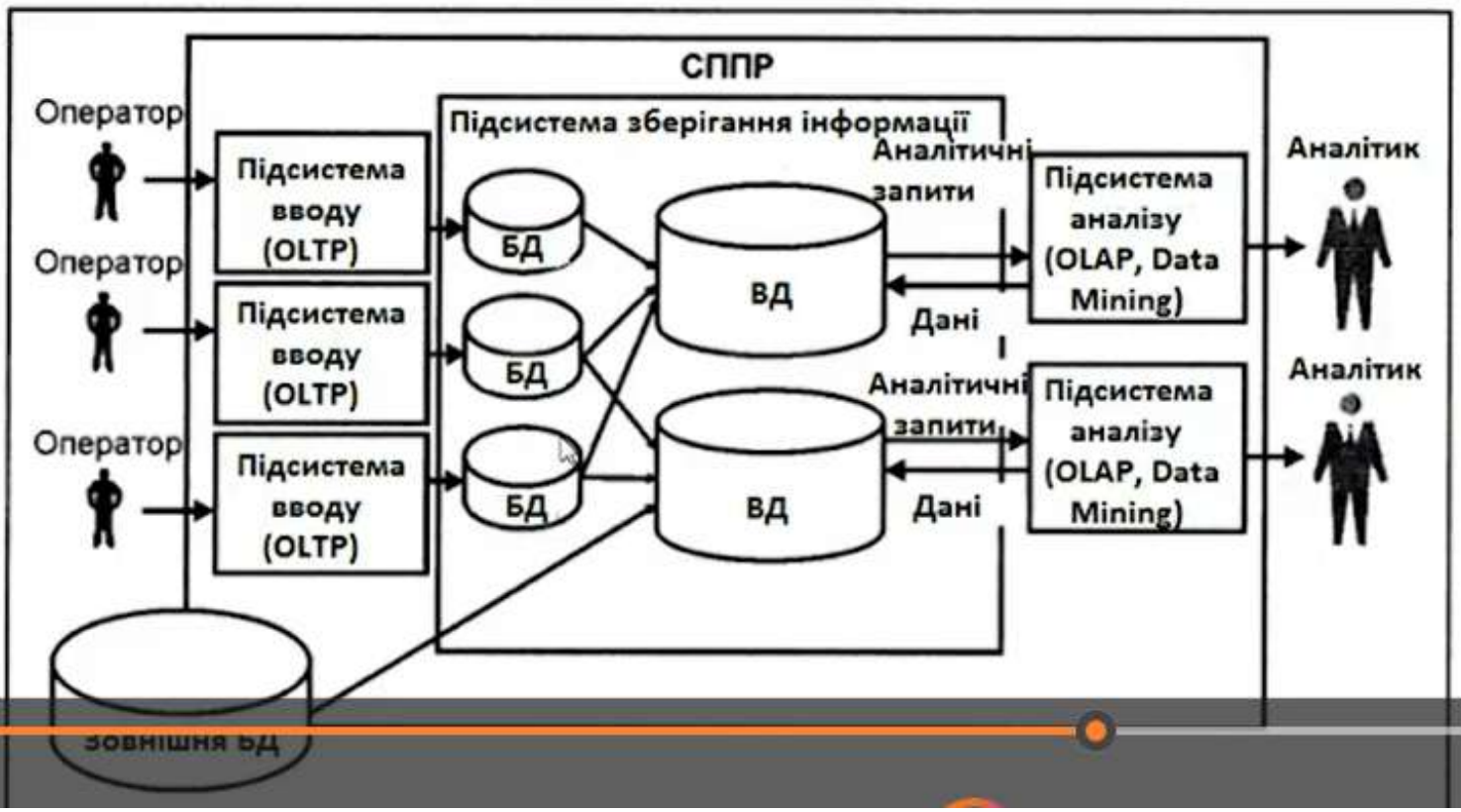


## Основні проблеми створення СД

- Необхідність інтеграції даних із неоднорідних джерел в розподіленому середовищі.
- Потреба ефективного зберігання і обробки великих об’ємів інформації.
- Необхідність багаторівневих довідників метаданих.
- Підвищення вимог до безпеки даних.

# Вітрина Даних (Data Mart )

- тематично об’єднані дані.



# Плюси і мінуси автономних Вітрин Даних (ВД)

+ Проектування ВД для відповідей на визначене коло запитань.
+ Швидке впровадження автономних ВД і отримання віддачі.
+ Спрощення процедур заповнення ВД і підвищення їх продуктивності за рахунок обліку потреб певного кола користувачів.

– Повторне використання даних.
– Проблеми, пов'язані з необхідністю підтримки несуперечливості даних.
– Відсутність загального вигляду ситуації.
– Додаткові затрати на розробку.

# Плюси і мінуси поєднання СД і ВД в одній системі

+ Простота створення і наповнення ВД, оскільки наповнення відбувається із єдиного стандартизованого надійного джерела очищених даних – з СД.
+ Простота розширення СППР за рахунок добавлення нових ВД.
+ Зниження навантаження на основне СД.

– Надлишковість (дані зберігаються як в СД, так і в ВД).
– Додаткові витрати на розробку СППР з СД і ВД.

Підводячи підсумок аналізу шляхів реалізації СППР з використанням концепції СД, можна виділити наступні архітектури таких систем:

- СППР з фізичним (класичним) СД;
- СППР з віртуальним СД;
- СППР з ВД;
- СППР з фізичним СД і з ВД.

У разі архітектури з фізичним СД і/або ВД необхідно приділити увагу питанням організації (архітектури) СД і переносу даних з ОДД в СД.

# Організація СД. Типи даних

Всі дані в СД діляться на основні категорії:

- детальні дані;
- агреговані дані;
- метадані;
- архівні дані.

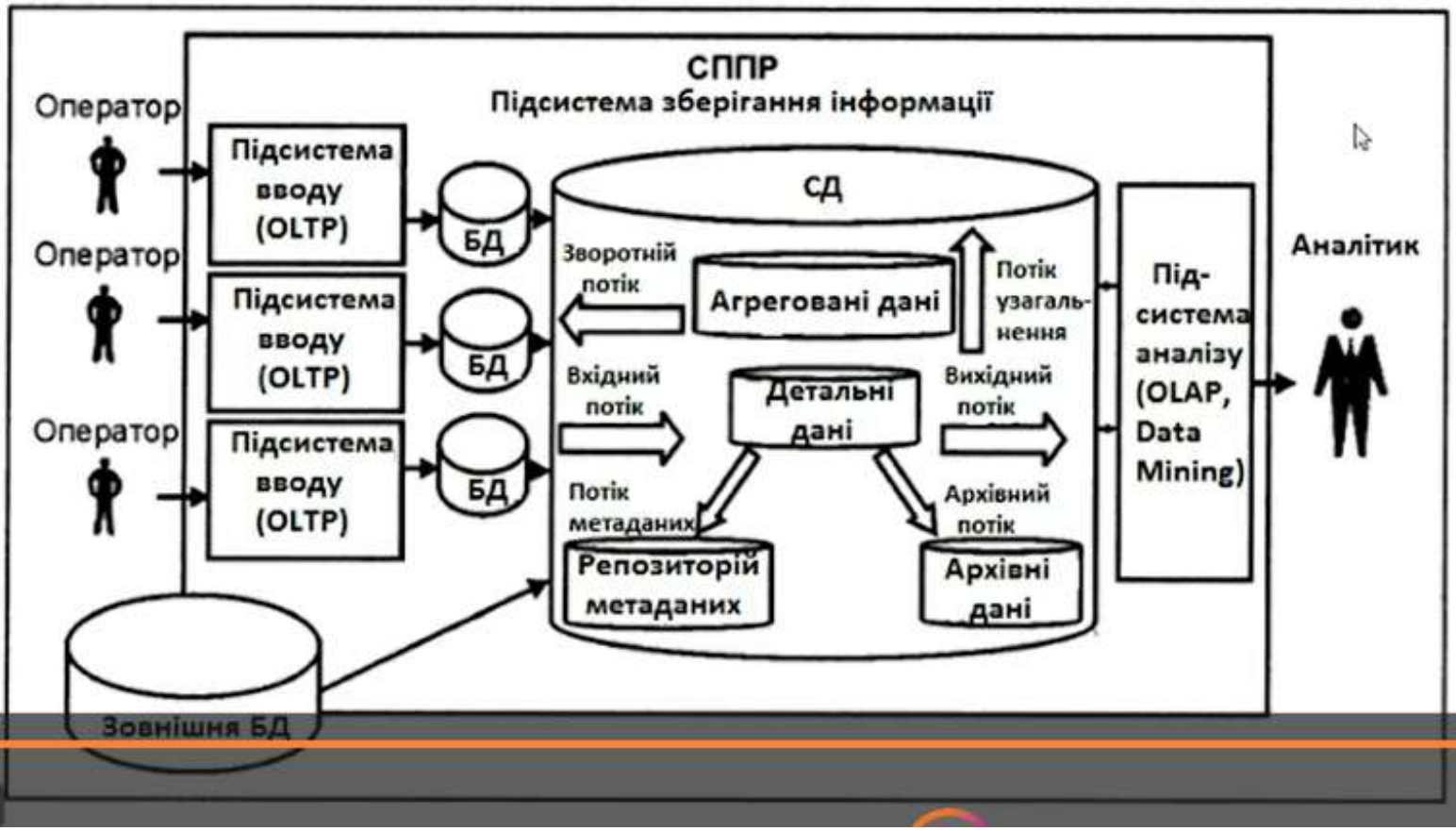
На основі детальних даних можна отримати агреговані (узагальнені) дані. Агрегування даних відбувається шляхом сумування числових фактичних даних за певними вимірами (просторами). В зв'язку з цим агреговані дані діляться на:

- адитивні – можна просумувати за всіма вимірами;
- напівадитивні – можна просумувати за деякими вимірами;
- неадитивні – не можна просумувати за жодним виміром.

## **Метадані** *(дані про дані).* **Згідно з концепцією Захмана (John A. Zachman) метадані повинні відповідати на такі питання – що, хто, де, як, коли і чому?**

- **Що? (Опис об'єктів)** Ці метадані описують об'єкти предметної області СД. Опис може містити: атрибути об'єктів, їх можливі значення, що ідентифікують поля в структурах даних, а також джерела відомостей про об'єкти і т.п.
- **Хто? (Опис користувачів)** Метадані, що відповідають на це питання містять профілі користувачів, які використовують дані: права доступу користувачів до даних, а також відомості про користувачів, які виконали операції над даними.
- **Де? (Опис місця зберігання)** Метадані описують місцезнаходження і взаємодію серверів, робочих станцій, джерел даних, а також розміщене на них програмне забезпечення і розподіл між ними даних.
- **Як? (Опис дій)** Ці метадані описують операції, що виконуються над даними. Описувані події могли виконуватися на різних етапах роботи з даними (перенесення з джерела даних, завантаження в сховище, вибірка зі сховища даних і т.п.).
- **Коли? (Опис часу)** Метадані, що відповідають на це питання описують момент або проміжок часу виконання різних операцій над даними.
- **Чому? (Опис причин)** Метадані цього типу описують причини виконання над даними операцій. Цими причинами може бути запит до даних, зміна кількості звернень до даних або досягнення певного значення контрольованого показника і т.п.

# Архітектура СД



**ETL–процес переносу даних включає в себе етапи видобування (E – extraction), перетворення (T – transformation) і завантаження (L – loading)**

1. Видобування даних

Видобування даних за допомогою засобів OLTP-систем в проміжні структури.

Видобування за допомогою допоміжних програмних засобів безпосередньо зі структур зберігання інформації (файлів, електронних таблиць, баз даних).

# ETL–процес (перетворення та завантаження)


Після того, як дані були перетворені для розміщення в СД, відбувається етап їх **завантаження**. В цьому процесі виконується запис перетворених детальних і агрегованих даних. Крім того при записі нових детальних даних частина старих може переміщатись в архів.

# ETL-процес. Проблеми очистки даних

Основні рівні проблеми очистки даних

1. Рівень комірки таблиці

- Орфографічні помилки
- Відсутність даних
- Фіктивні значення
- Логічно невірні значення
- Закодовані значення
- Складені значення

2. Рівень запису

Проблема суперечливості значень в різних полях

3. Рівень таблиці БД

- Порушення унікальності
- Відсутність стандартів

4. Рівень одиночної БД

Порушення цілісності даних

5. Рівень множини БД

Проблеми з неоднорідністю структур БД, а також інформації всередині цих баз

# ETL-процес. Реалізація очистки даних

# Тести

1. База даних – це:

а) модель деякої предметної області, яка складається із зв'язаних між собою даних про об'єкти, їх властивості та характеристики;

б) централізоване сховище даних, що забезпечує зберігання, доступ, первинне опрацювання і пошук інформації;

в) система мовних, алгоритмічних, програмних, технічних і організаційних засобів підтримки інтегрованої сукупності даних, а також самі ці дані;

г) предметно-орієнтований, інтегрований, незмінний набір даних, такий що підтримує хронологію і організований для цілей підтримки прийняття рішень.

# Тести

2. Сховище даних – це:

а) сукупність екземплярів різних типів записів і відношень між записами і елементами;

б) централізоване сховище даних, що забезпечує зберігання, доступ, первинне опрацювання і пошук інформації;

в) система мовних, алгоритмічних, програмних, технічних і організаційних засобів підтримки інтегрованої сукупності даних, а також самі ці дані;

г) предметно-орієнтований, інтегрований, незмінний набір даних, такий що підтримує хронологію і організований для цілей підтримки прийняття рішень.

# Тести

3. Вимоги до сховища даних – це:

а) предметна орієнтація;

б) змінюваність в часі;

в) децентралізованість;

г) кросплатформність.

4. Класифікаційна схема Захмана описує:

а) детальні дані;

б) агреговані дані;

в) метадані.

# Тести

5. Відмінності сховищ даних від реляційних баз даних:

а) отримання інформації з різних джерел даних;

б) наявність метаданих;

г) підтримка третьої нормальної форми;

д) орієнтованість даних на аналітичне опрацювання.

# Тести

6. Вітрина даних – це:

а) зріз сховища даних, масив тематичної, вузьконапрямленої інформації, що орієнтований на користувачів однієї робочої групи;

б) предметно-орієнтований, інтегрований, незмінний набір даних, такий що підтримує хронологію і організований для цілей підтримки прийняття рішень;

в) єдине сховище даних проблемної області, що забезпечує проблеми всіх її функціональних частин, які мають потребу в засобах аналізу даних;

г) предметно-орієнтований, інтегрований, змінюваний набір даних, який містить поточну деталізовану інформацію.