
Welcome to course

**COMP810: Data Warehousing and
Big Data**

Outlines

- Course information
- Teaching team
- Learning outcomes
- Course contents and schedule
- Assessments
- Resources
- Lecture 01



Course Information

| | |
|-----------------------|--------------------------------------|
| Course Title: | Data Warehousing and Big Data |
| Course Code: | COMP810 |
| Prerequisites: | None |
| Co-requisites: | None |
| Level: | 8 |

Teaching Team

- **Dr Muhammad Asif Naeem**
Paper leader & lecturer
Email: mnaeem@aut.ac.nz
Office: WZ336A
Ext: 5083
- **Amr Mahmoud**
TA
Email : amr.mahmoud@aut.ac.nz



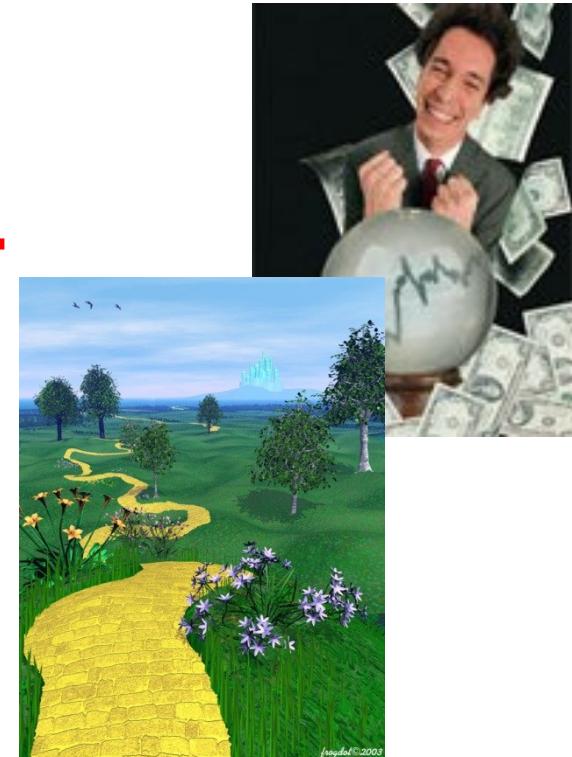
Why should you be here?

- Bad decisions can lead to disaster
 - Data Warehousing is at
the base of decision support systems



Why should you be here?

- Data Warehousing & OLAP are important
- They help to
 - Understand the information **hidden** within an organization's data
 - See data from different angles:
product, client, time, geographical area
 - Get adequate statistics to get your point of argumentation across
 - Get a glimpse of the future...



Why should you be here?

- And because you **love databases...**

The screenshot shows a job listing on the seek.com.au website. The top navigation bar includes the seek logo, a sign-in/register link, and tabs for Job Search, Profile, and My Activity. Below the navigation is a back-link to search results. The main content area displays a job listing for a "Data Warehouse Support Consultant (WhereScape RED) - Urgent" position. The listing is from "absoluteIT RECRUITMENT SPECIALISTS". It specifies the location as Auckland Central, a salary of \$105,000, and a full-time work type. A red circle highlights the salary information. Below the listing, there's a section titled "The essential skills that will encourage success in your application are:" followed by a bulleted list of required skills.

seek

Sign in or Register

Job Search Profile My Activity

« Back to search results

Data Warehouse Support Consultant
(WhereScape RED) - Urgent

Absolute IT- 96% of job seekers we place would recommend us to others* - [More jobs by this advertiser](#)

absoluteIT
RECRUITMENT SPECIALISTS

Location: Auckland ▶ Auckland Central

Salary: \$105,000

Work type: Full Time

The essential skills that will encourage success in your application are:

- **Datawarehouse Tools** - a minimum of 4 years commercial support of either; MS Stack, Teradata or Oracle based technologies and tools.
- **ETL Specific** - exposure to WhereScape RED is essential.
- **Functions** - experience in dealing with customers via both phone and on a face to face basis (Database / Enterprise Data Warehouse support).

Course Contents and Schedule

| WEEK | LECTURE | LAB |
|---------------------------|--|--|
| 1 | Introduction to course, Introduction to Data Warehouse | Writing Basic SQL Statements; |
| 2 | Assessment 1 handed out and academic writing | SQL Operators – Restricting and Sorting Data |
| 3 | DW Life Cycle and Basic Architecture | Retrieving Data from Multiple tables |
| 4 | DW Architecture (Continued) | SQL functions and Aggregating data using Group Functions |
| 5 | Logical Model | Table Creation, Alteration and Applying constraints |
| 6 | Physical Model | DML – Insert, Update and Delete Data |
| 7 | Indexes Assessment 2 handed out | SQL Practice |
| Mid Semester Break | | |
| 8 | Optimizations | Creating of DW Schema |
| 9 | OLAP Operations and Queries | PLSQL datatypes & control statements |
| 10 | Building the DW | PLSQL collections & records and static SQL |
| 11 | Big Data | Dynamic SQL and PLSQL subprograms |
| 12 | Big Data Continued | No Lab |

Assessments

| Assessment type | Date |
|-----------------------|---------|
| Research Report (40%) | Week 7 |
| Project (60%) | Week 13 |



Important: To pass the paper, student needs to attempt both assessments and obtain a C- grade overall.

Recommended Literature

- Building the Data Warehouse
 - William H. Inmon
 - Wiley, ISBN 978-0-7645-9944-6
- The Data Warehouse Toolkit
 - Ralph Kimball & Margy Ross
 - Wiley, ISBN 978-1-118-53080-1
- Big Data
 - Nathan Marz, James Warren
 - Wiley, ISBN 978-1-617290-34-3

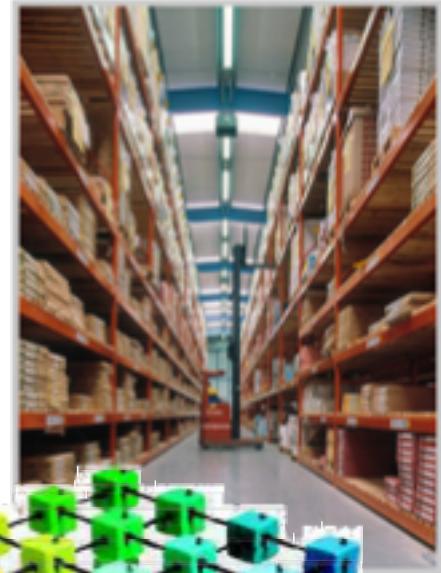


Lecture 01

Introduction to Data Warehouse

This week

1. What is a data warehouse?
2. Applications and users



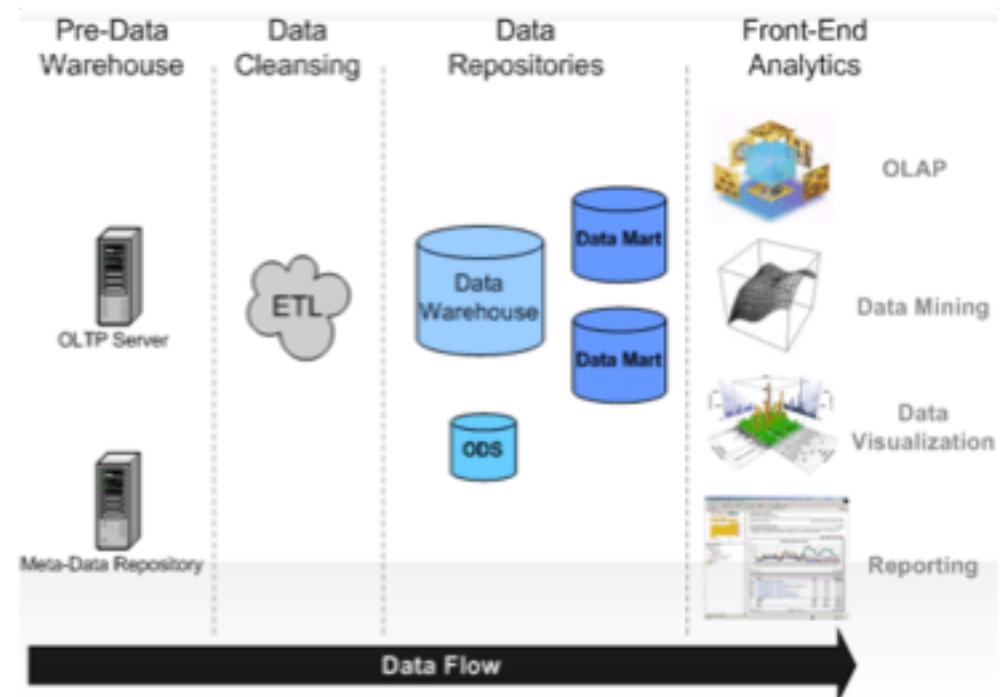
What is a data warehouse?

- Basically a **very large database**...
 - Not all **very large databases** are data warehouses, but all data warehouses are pretty large databases
 - Nowadays a warehouse is considered to start at around 800 GB and goes up to several TB
 - It spans over several servers and needs an impressive amount of computing power



What is a data warehouse?

- More specific, a **collective data repository**
 - Containing snapshots of the operational data (history)
 - Obtained through data cleansing (Extract- Transform- Load)
 - Useful for analytics



What is a data warehouse?

- Compared to other solutions it...
 - Is suitable for **tactical/strategic focus**
 - Implies a **small number of transactions**
 - Implies **large transactions** spanning over a long period of time

| | OLTP | ODS | OLAP | DM / DW |
|-------------------|----------------------|------------------------|---------------|----------------------|
| Business Focus | Operational | Operational / Tactical | Tactical | Tactical / Strategic |
| End User Tools | Client/Server or Web | Client/Server or Web | Client/Server | Client/Server or Web |
| DB Technology | Relational | Relational | Cubic | Relational |
| Transaction Count | Large | Medium | Small | Small |
| Transaction Size | Small | Medium | Medium | Large |
| Transaction Time | Short | Medium | Medium | Long |
| DB Size in GB | 10–400 | 100–800 | 100–800 | 800–80,000 |
| Data Modeling | Traditional ERD | Traditional ERD | N/A | Dimensional |
| Normalization | 3–5 NF ¹ | 3 NF | N/A | 0 NF |

Some Definitions

- Experts say...
 - **Ralph Kimball:** “a copy of transaction data specifically structured for query and analysis”
 - **Bill Inmon:** “A data warehouse is a:
 - Subject oriented
 - Integrated
 - Non-volatile
 - Time variantcollection of data in support of management’s decisions.”



Inmon Definition (cont'd.)

- **Subject oriented**

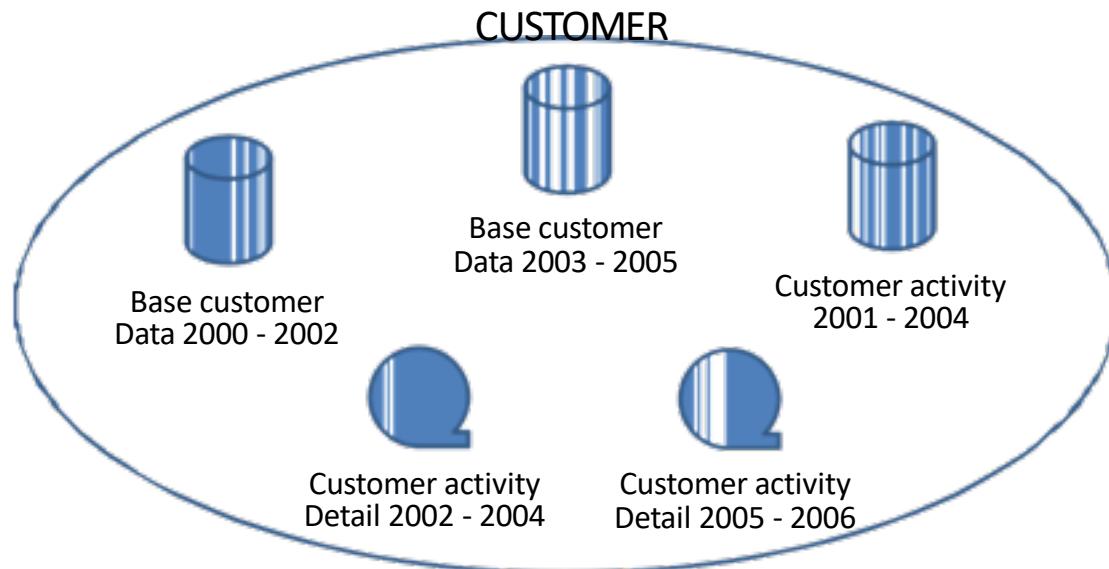
- The data in the data warehouse is organized so that all the data elements relating to the same real-world event or object are **linked together**

- Typical subject areas in DWs are Customer, Product, Order, Claim, Account,...



Inmon Definition (cont'd.)

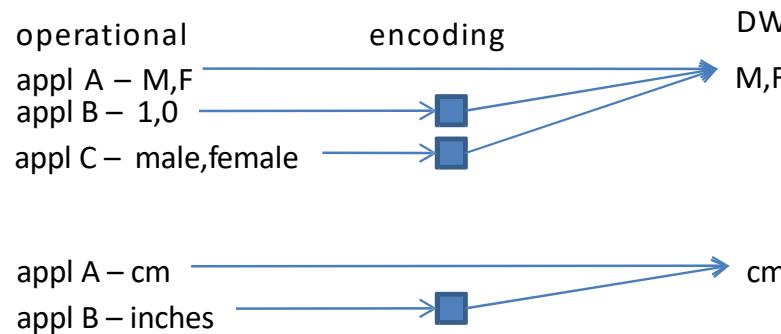
- **Subject oriented**
 - Example: customer as subject in a DW
 - DW is organized in this case by the customer
 - It may consist of 10, 100 or more physical tables, all related



Inmon Definition (cont'd.)

- **Integrated**

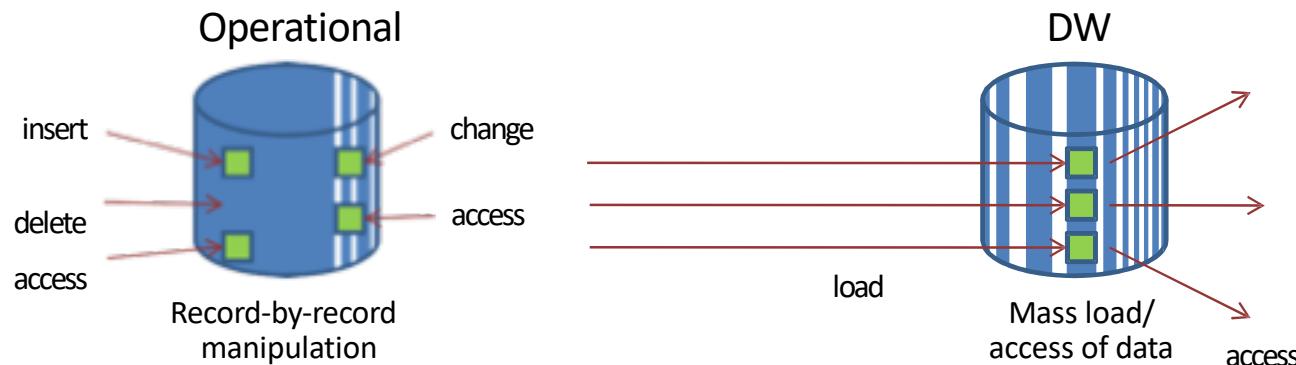
- The data warehouse contains data from **most or all** of an organization's operational systems and this data is made **consistent**
- E.g. gender, measurement, conflicting keys, consistency,...



Inmon Definition (cont'd.)

- **Non-volatile**

- Data in the data warehouse is **never over-written** or **deleted** - once committed, the data is static, read-only, and retained for future reporting
- Data is loaded, but not updated
- When subsequent changes occur, a new snapshot record is written



Inmon Definition (cont'd.)

- **Time-varying**
 - The changes to the data in the data warehouse are tracked and recorded so that reports can be produced showing **changes over time**
 - Different environments have different time horizons
- associated
 - While for operational systems a 60-to-90 day time horizon is normal, data warehouse has a
 - 5-to-10 year horizon



General Definition

- More general, a DW is a
 - Repository of an organization's electronically stored data
 - Designed to facilitate reporting and analysis



Typical Features

- DW typically...
 - Reside on computers **dedicated to this function**
 - **Run on DBMS** such as Oracle, IBM DB2, Teradata or Microsoft SQL Server
 - Retain data for **long periods of time**
 - **Consolidate data** obtained from a variety of sources
 - Are built around their own **carefully designed data model**

Use case

- DW stands for big data volume, so lets take an example of **2 big companies**, a retailer, say Walmart and a RDBMS vendor, Sybase:
 - Walmart CEO: *I want to keep track of sales in all my stores simultaneously*
 - Sybase consultant: *You need our wonderful RDBMS software. You can stuff data in as sales are rung up at cash registers and simultaneously query data right in your office*
 - So Walmart buys a \$1 milion Sun EI0000 multi-CPU server, a \$500 000 Sybase license, a book “Database Design for Smarties”, and build themselves a normalized SQL data model



Use case (cont'd.)

- After a few months of stuffing data into the tables...
a Walmart executive asks...
 - *I have noticed that there was a Colgate promotion recently, directed to people who live in small towns. How much toothpaste did we sell in those towns yesterday?*
 - Translation to a query:

```
select sum(sales.quantity_sold) from sales,products,product_categories,manufacturers,  
stores,cities where manufacturer_name = 'Colgate'  
and product_category_name = 'toothpaste' and cities.population < 40 000  
and trunc(sales.date_time_of_sale) = trunc(sysdate-1) and sales.product_id =  
products.product_id  
and sales.store_id = stores.store_id  
and products.product_category_id = product_categories.product_category_id and  
products.manufacturer_id = manufacturers.manufacturer_id  
and stores.city_id = cities.city_id
```

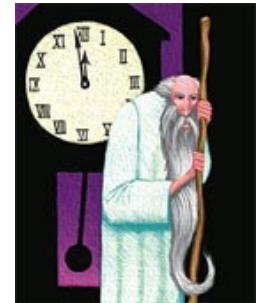


Use case (cont'd.)

- The tables contain large volumes of data and the query implies a **6 way join** so it will take some time to execute
- The tables are at the **same time also updated** by new sales
- Soon after executive start their quest for marketing information store employees notice that there are times during the day when it is impossible to process a sale



Any attempt to update the database results in freezing the computer up for 20 minutes

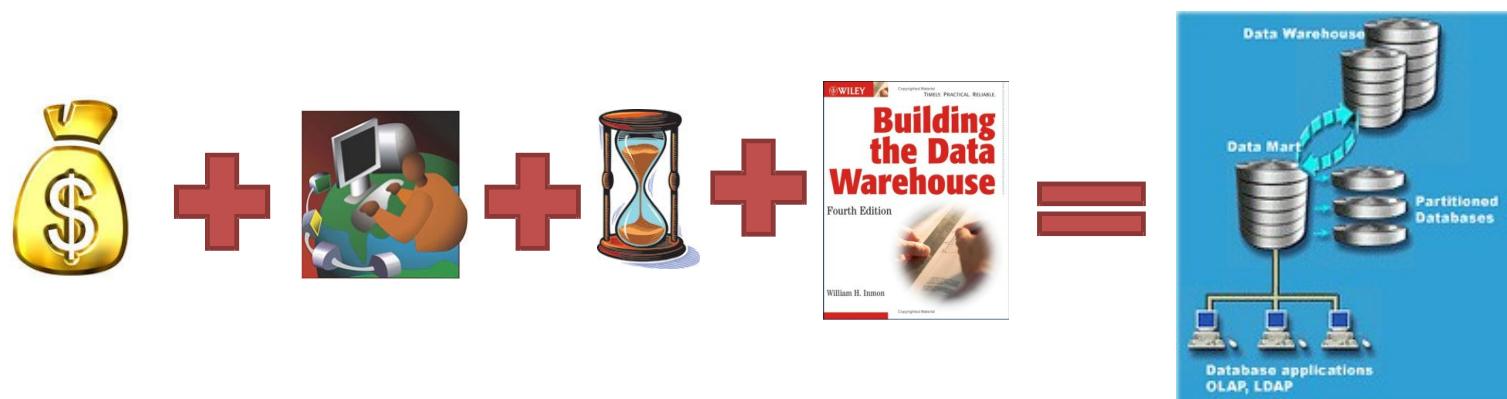


Use case (cont'd.)

- In minutes...the Walmart CEO calls Sybase tech support
 - **Walmart CEO:** WE TYPE IN THE TOOTHPASTE QUERY AND OUR SYSTEM HANGS!!!
 - **Sybase support:** Of course it does! You built an **on-line transaction processing (OLTP)** system. You can't feed it a **decision support system (DSS)** query and expect things to work!
 - **Walmart CEO:** !@%\$#. I thought this was the whole point of SQL and your RDBMS...to query and insert simultaneously!!
 - **Sybase support:** Uh, not exactly. If you're **reading** from the database, nobody can **write** to the database. If you're **writing** to the database, nobody can **read** from the database. So if you've got a query that takes 20 minutes to run and don't specify **special locking instructions**, nobody can update those tables for 20 minutes.

Use case (cont'd.)

- Walmart CEO: *It sounds like a bug.*
- Sybase support: *Actually it is a feature. We call it **pessimistic locking**.*
- Walmart CEO: *Can you fix your system so that it doesn't lock up???*
- Sybase support: *No. But we made this great loader tool so that you can copy everything from your **OLTP system** into a separate **Data Warehouse system** at 100 GB/hour*
- After a while...



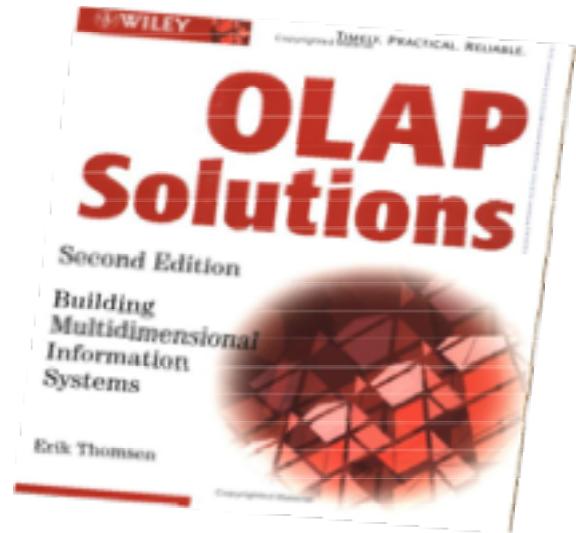
OLTP vs. DW

- OLTP (**O**n**L**ine **T**ransaction **P**rocessing):
 - Also known under the name of **operational data**, it represents day-to-day operational business activities:
 - Purchasing, sales, production distribution, ...
 - Typically for **data entry** and **retrieval** transaction processing
 - Reflects only the **current state** of the data



OLTP vs. DW (cont'd.)

- OLAP (OnLine Analytical Processing):
 - Represents **front-end analytics** based on a DW repository
 - It provides information for activities like
 - Resource planning, capital budgeting, marketing initiatives,...
 - It is **decision oriented**



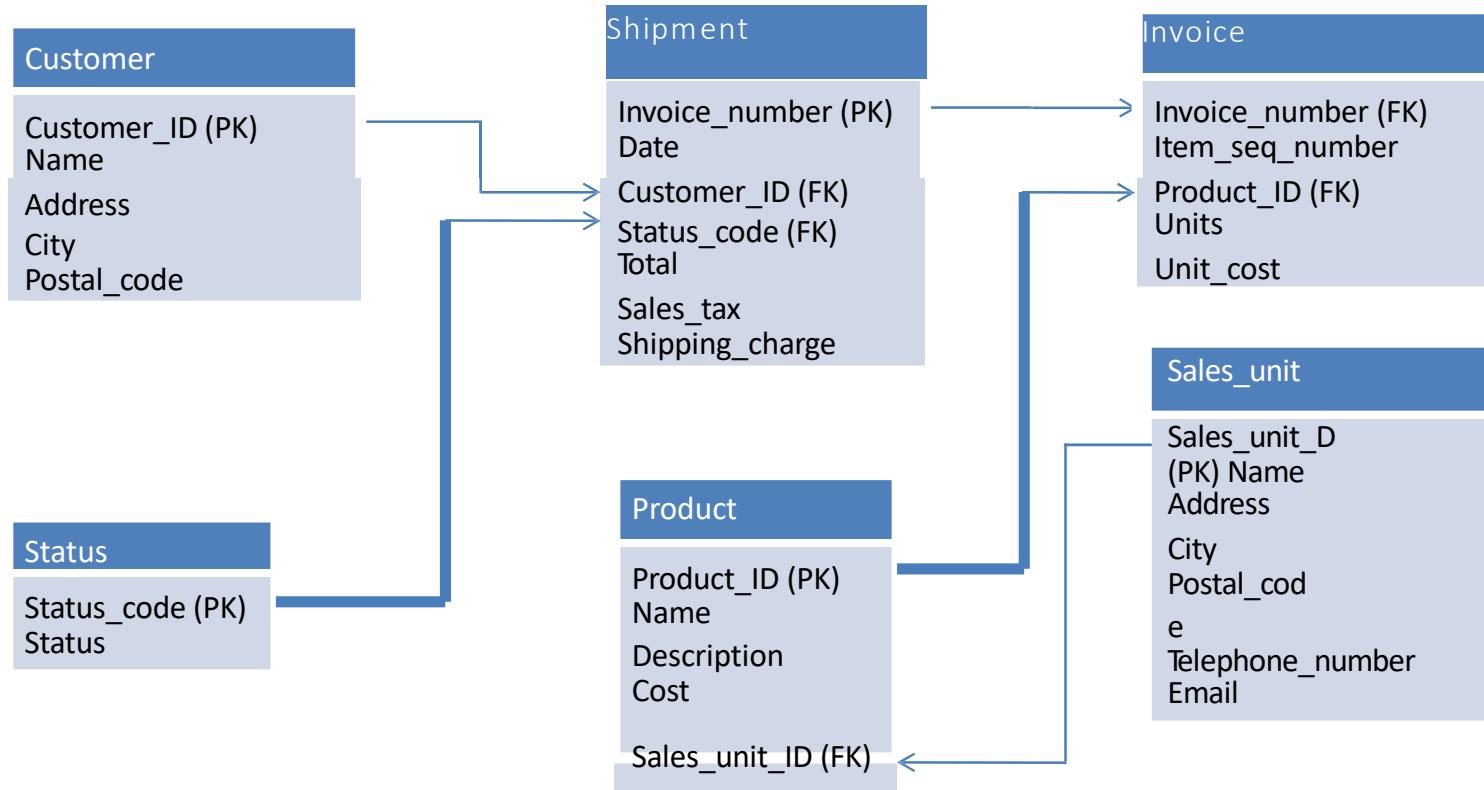
OLTP vs. DW (cont'd.)

- Properties

| Operational DB | DW |
|-------------------------|--------------------------|
| Mostly updates | Mostly reads |
| Many small transactions | Queries long, complex |
| MB-TB of data | GB-PB of data |
| Raw data | Summarized data |
| Clerical users | Decision makers |
| Up-to-date data | May be slightly outdated |

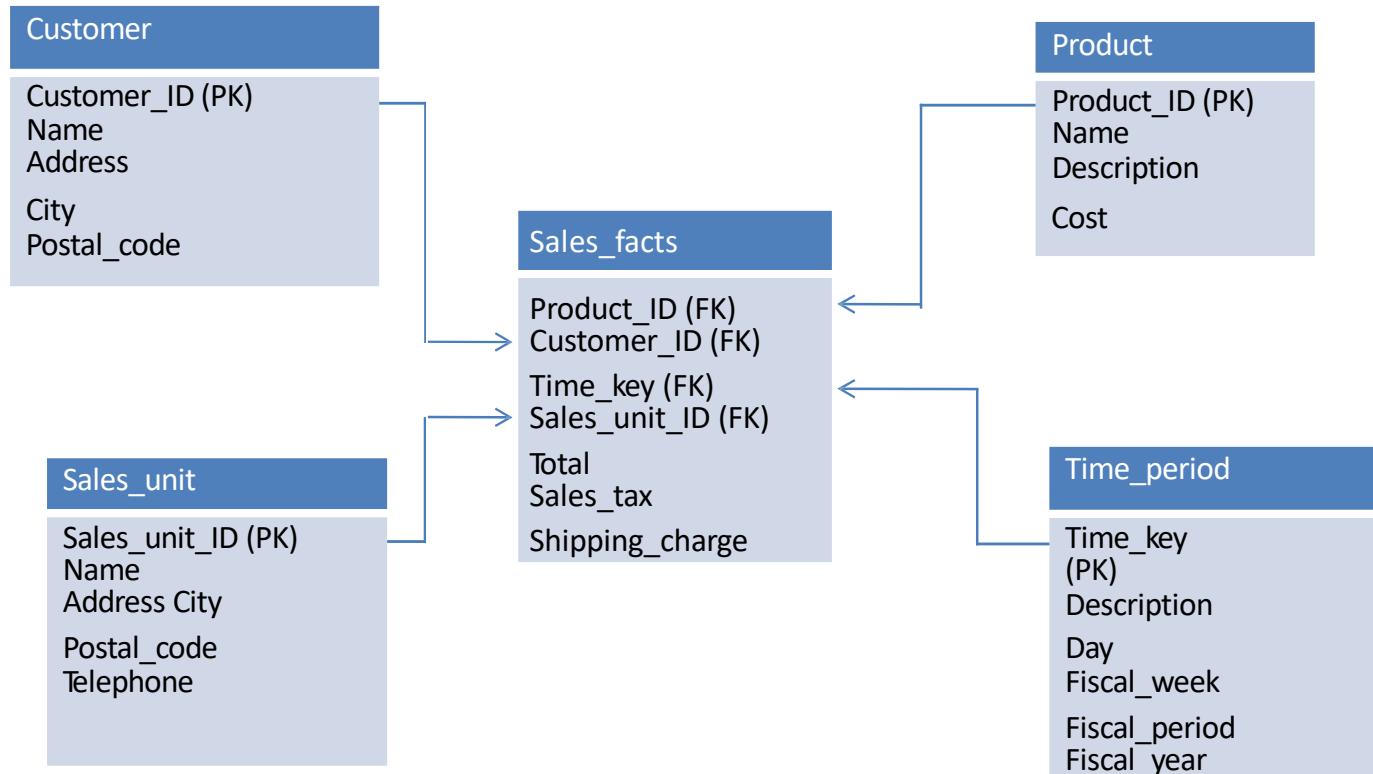
OLTP vs. DW (cont'd.)

- Consider a **normalized database** for a store
 - The tables would look like this...



OLTP vs. DW (cont'd.)

- If we were to set up a **DW** for that store, we would start by building the following **star schema**

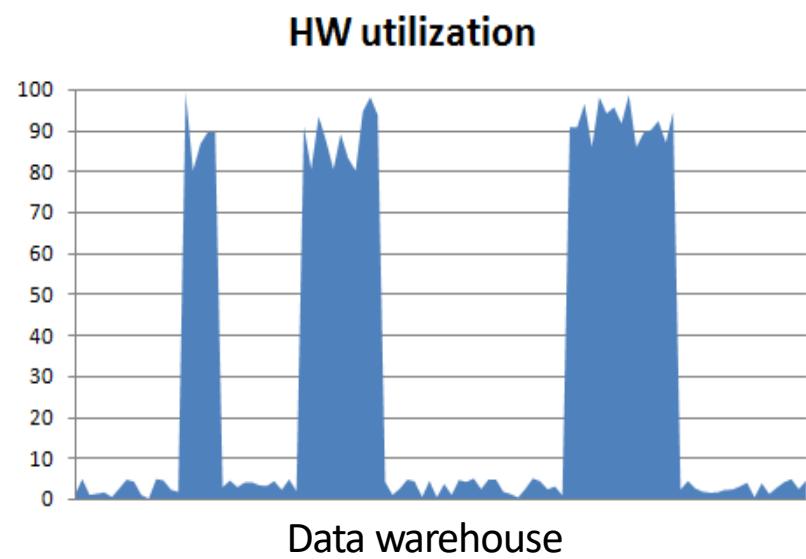
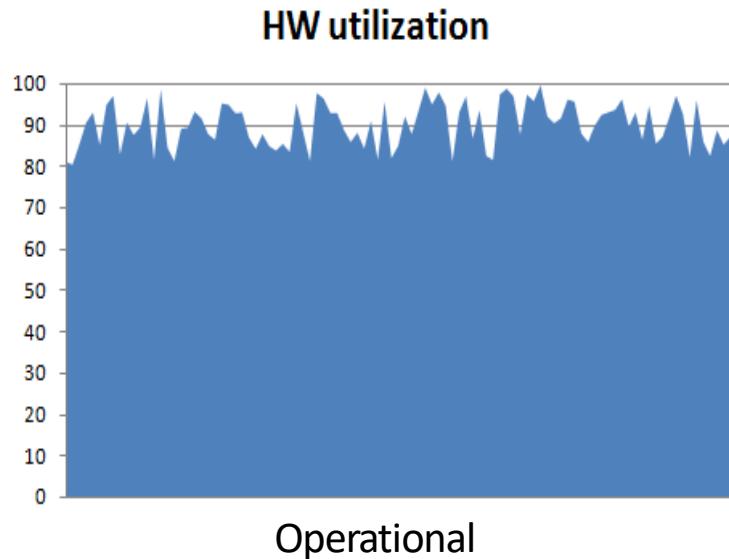


OLTP vs. DW (cont'd.)

- **Basic insights** from comparing OLTP and DWs
 - A DW is a separate (**RDBMS**) installation that contains copies of data from on-line systems
 - Physically separate hardware may not be absolutely necessary if you have lots of **extra computing power**, but it is recommended
 - With an **optimistic locking** DBMS you might even be able to get away for a while with keeping just one copy of your data

OLTP vs. DW (cont'd.)

- There is an essentially different pattern of **hardware utilization** between on-line and analytical processing

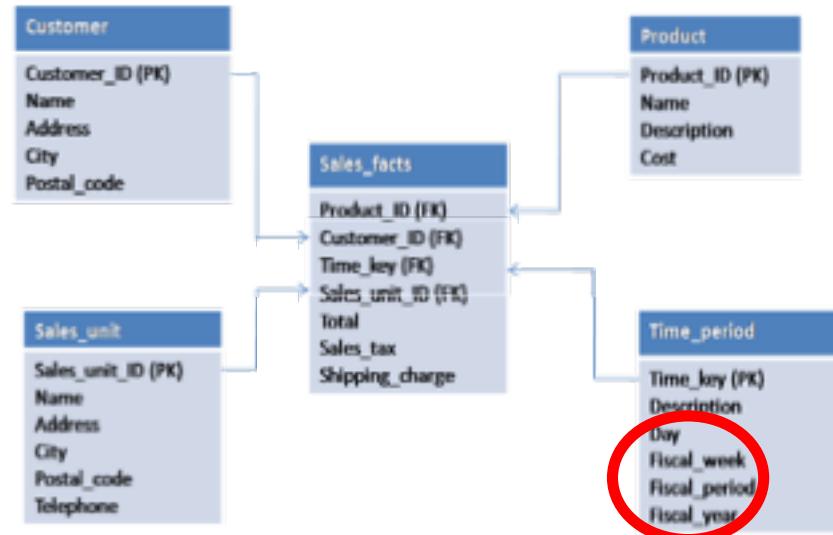


Applications of DW

- **Typical questions** which can be answered with DW & OLAP
 - How much did sales unit A earn in January?
 - How much did sales unit B earn in February?
What was their combined sales amount for the first quarter?
- Answering these questions with **SQL-queries** is difficult
 - Complex query formulation necessary
 - Process is likely to be slow due to complex joins and multiple scans

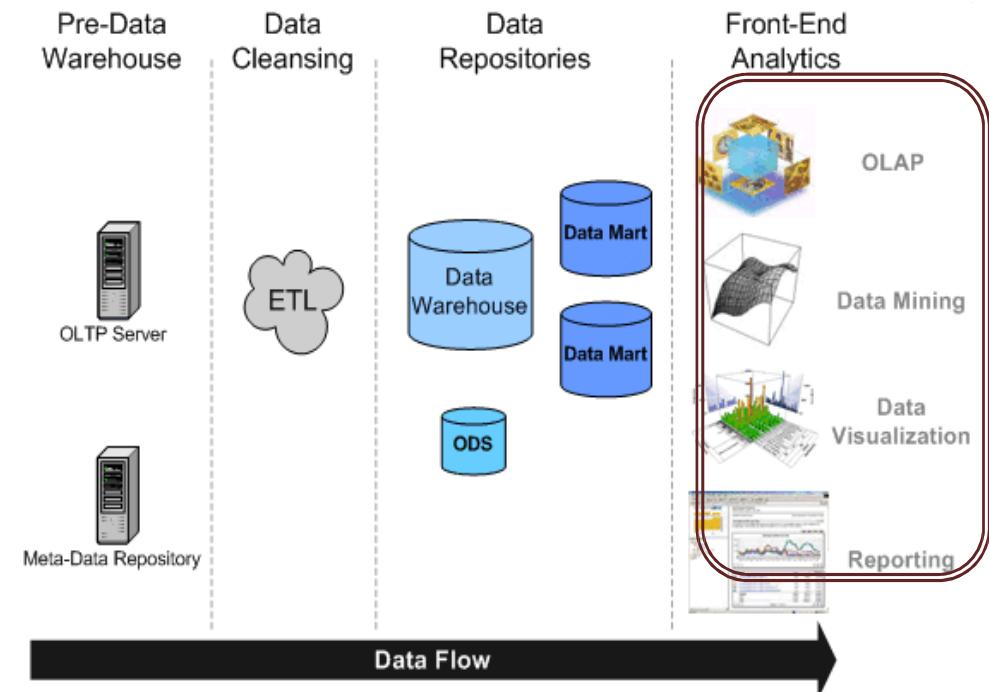
Applications of DW (cont'd.)

- Why such questions can be answered better with a DW?
 - Because in a DW tables are **rearranged and pre-aggregated** (known as *computing cubes*)
e.g. Years, weeks, days etc.
 - The tables arrangement is subject oriented, usually some **star schema**



Applications of DW (cont'd.)

- A DW is the base repository for **front-end analytics**
 - OLAP
 - KDD (**Knowledge Discovery in Databases**) a data mining process
 - Data visualization
 - Reporting



Applications of DW (cont'd.)

- OLAP is a form of information processing and thus needs to provide **timely, accurate** and **understandable** information
 - timely is however a relative term:
 - In OLTP we expect an update to go through in a matter of **seconds**
 - In OLAP the time to answer a query can take **minutes, hours or even longer**
- There are many **flavors** of OLAP
 - ROLAP, DOLAP, MOLAP, WOLAP, HOLAP, ...



Applications of DW (cont'd.)

- **KDD** (Data Mining)
 - Constructs **m odel s** of the data in question
 - Models can be viewed as high level summaries of the underlying data

| ID | Name | ZIP | Sex | Age | Income | Children | Car | Spent |
|-----|---------|-------|-----|-----|----------|----------|----------|----------|
| 12 | Peter | 38106 | M | 35 | € 55,000 | 2 | Mini Van | € 210.00 |
| 15 | Gabriel | 38100 | M | 32 | € 56,000 | 0 | SUV | € 30.00 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 122 | Claire | 38106 | F | 21 | € 42,000 | 0 | Coupe | € 50.00 |

Applications of DW (cont'd.)

- Based on this example a query returns the **data** that fulfills the constraints
 - `SELECT * FROM CUSTOMER_TABLE WHERE TOTAL_SPENT > €100;`
- Data mining might return the following **set of rules** for customers spending more than €100:
 - `IF AGE > 35 AND CAR = 'MINIVAN' THEN TOTAL SPENT > €100`
 - `IF SEX = 'M' AND ZIP = 38106 THEN TOTAL SPENT > €100`

Applications of DW (cont'd.)

- It answers **questions** like
 - Which products or customers are more profitable
 - Which outlets have sold the least this year
- In consequence it motivates **decisions** like
 - Which products should have their production increased
 - Which customers should be targeted for special promotions
 - Which outlets should be closed



Who is the user?

- Users of DW are called **DSS analysts** and usually are business persons
 - Their primary job is to **define** and **discover** information used in corporate **decision-making**
 - The way they think
 - “Give me what I say I want, and then I can tell you what I really want”
 - They work in explorative manner



Who is the user? (cont'd.)

- Typical **explorative** line of work
 - “Ah! Now that I see what the possibilities are, I can tell what I really want to see. But until I know what the possibilities are, I cannot describe exactly what I want...”
- This usage has fundamental effect on **the way a DW is developed**
 - The classical **system development life cycle** assumes that the requirements are known at the start of design
 - The DSS analyst starts with existing requirements, but **factoring in new requirements** with time

Summary

Summary

- Data Warehouse - Introduction
 - DW Definitions
 - OLTP vs. DW
 - Applications of DW

Next lecture

- Academic Writing
- Briefing about Assessment I