GROUP 8

SQL

- → Structured query language
- → a programming language for storing and processing information in a relational database.
- → A relational database stores information in tabular form, with rows and columns representing different data attributes and the various relationships between the data values

History

- → 1970 – IBM engineers Raymond Boyce and Donald Chamberlin created the SQL programming language.
- → Following the publication of Edgar Frank Codd's paper, "A Relational Model of Data for Large Shared Data Banks, " in 1970, the programming language, then known as SEQUEL, was developed

Benefits of SQL

- → SQL is portable
  - It runs on local internet and intranet systems.
- → Processes queries quickly
  - No matter how large data might be, SQL can retrieve it quickly and efficiently.
  - It can also achieve processes such as insertion, deletion and data manipulation relatively quickly.
- → Doesn't require coding skills
  - Coding is a complicated way of communicating with computers. Also called computer programming, coding can require lots of practice and knowledge before use, making it difficult for others to interpret.
- → Standardized language
  - The standardized language used in SQL makes it highly accessible to all users.
- → Open source code
  - MySQL, MariaDB and PostGres offer free SQL databases

SQL Commands

- → Data Definition Language (DDL)
  - Used to update or manipulate a database structure. Commands include CREATE, ALTER, DROP and RENAME.
- → Data Query Language (DQL)
  - Used to fetch data from the database. It only uses the SELECT command.
- → Data Manipulation Language (DML)
  - Enables the modification of a database. Commands include INSERT, UPDATE and DELETE.
- → Data Control Language (DCL)
  - Used to set privilege and permission parameters within the database structure. Commands include GRANT and REVOKE.
- → Transaction Control Language (TCL)
  - Used to manage changes made by DML. It enables these changes to be grouped into logical transactions. Commands include COMMIT, ROLLBACK and SAVEPOINT

Advantages of SQL

- → Multiple Data Views
- → Interactive Language
- → Backup and Recovery
- → Data Integrity
- → Data Consistency

Disadvantages of SQL

- → Poor interface
- → Cost inefficient
- → Partial Control
- → Security

Relationship in SQL

- → One-to-one relationship
- → One-to-many relationship
- → Many-to-many relationship

SQL Technologies

- → SQL in databases
  - Relational Databases
  - Data Warehouses and Analytical Databases
- → Data Processing Technologies
  - Distributed Data Processing Technologies
  - Full-text search engines
- → SQL in Everyday Apps
  - Spreadsheets
  - Smartphone apps
  - Web Browsers

Application of SQL

- → In Healthcare
  - make use of SQL to analyze large data sets containing information about patient behavior, their medical conditions, and demographics.
- → In the finance industry
  - complicated database systems powered by SQL which enable the extraction of actionable insights that serve to check for fraud and enable the delivery of personalized experiences to users.
- → In database administration
  - make use of SQL to capture and process confidential information about users, employees, students, or patients, without much hassle.
- → In data analysis

- The use of SQL can streamline the process of gaining insights from huge volumes of data, utilizing a variety of conditional commands.

→ In social media
- make use of SQL to store the profile information of users and allows them to update the database of their app when users create new posts or share photos, and it also facilitates the recording of messages, enabling users to retrieve messages later.

→ In data science
- make use of SQL code and algorithms to create a data model that analysts can use to explore data and discover business-specific trends and combinations in that data.

GROUP 9 OODBMS

OODBMS History

→ 1960 – 1970
- Ole-Johan Dahl and Kirsten Nygaard - The Simula Programming Language

→ 1980
- ACM SIGMOD Conference, J. Rosenberg - "GemStone" Project

→ 1990
- CAD/CAM, telecommunications, and scientific research

→ Present
- Object-Relational, XML, and NoSQL databases

Major Components

→ Object structure
- The structure of an object refers to the properties that an object is made up of.

→ Object classes
- An object which is a real-world entity is an instance of a class.

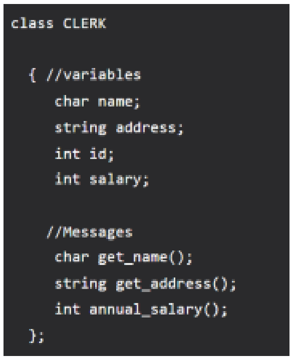
→ Object identity
- Distinct value assigned to each object by the system.

Object structure 3 types of components

→ Messages
- Acts as a communication

→ Methods
- Body of code that is executed

→ Variables
- It stores the data of an object



```
class CLERK

  { //variables
     char name;
     string address;
     int id;
     int salary;

    //Messages
     char get_name();
     string get_address();
     int annual_salary();
  };
```

Features of OODBMS

→ Object-Oriented Data Model
→ Complex Data Type
→ Automatic schema management
→ High performance
→ Data integrity
→ Concurrency control
→ Scalability
→ Support for transactions

Advantages of OODBMS

→ Supports Complex Data Structures
→ Improved Performance
→ Reduced Development Time
→ Supports Rich Data Types
→ Scalability

Disadvantages of OODBMS

→ Limited Adoption
→ Lack of Standardization
→ High Costing
→ Integration of other systems
→ Scalability Challenges

Benefits using OODBMS

→ Enterprise Applications
→ Real-time Data Processing
→ Object-oriented programming languages
→ Geospatial data
→ Multimedia data

Applications of OODBMS

- Commonly used in applications that require high performance, calculations, and faster results.
- Some of common applications that use object databases are real-time systems, architectural & engineering for 3d modeling, telecommunications, and scientific products, molecular science, and astronomy

→ Geographic information systems (GIS)
- manage geographical data, such as maps, spatial objects, and their attributes.

  - The hierarchical and complex relationships in GIS data are well - supported by OODBs.
- → Computer-aided design (CAD) systems
  - Cad applications often involve intricate 2d or 3d models with complex interconnections and relationships between objects.
  - Allow for efficient querying and manipulation
- → Network management
  - To represent network devices, configurations, and performance data.
- → E-commerce and product catalogs
  - To manage a vast array of product information with attributes and hierarchies.
  - Can help organize and maintain product catalogs efficiently
- → Data warehousing
  - For managing and analyzing large and complex datasets.
  - Can represent data structures with complex relationships and provide an efficient way to query and analyze the data

Diagram Representations

- → Polymorphism
  - the capability of an object to take multiple forms
- → Inheritance
  - the principle that a class can base itself on properties of another class
- → Encapsulation
  - an object contains both the data structures and the methods to manipulate the data structures
- → abstraction
  - Capturing necessary information and reduces complexity

GROUP 10

Web technologies and DBMS

Overview/Concepts

- → Programming languages
- → Web frameworks
- → Databases

Benefits of web technologies and DBMS

- → Security
  - encryption and password protection
  - impossible for outsider to access database
- → Accessibility
  - access data from any internet enabled device
  - never worry about losing valuable data because it is stored on another device
- → Reliability and scalability
  - accessed by multiple users simultaneously
  - easier to handle simultaneous requests without slowing down or crashing
- → Ease of maintenance for IT Staff
  - problems can be isolated and quickly fixed
  - reduces downtime for users; reduce cost for IT staff responsible for maintenance
  - database automation tools make tasks easier and safer

Advantages

- → accessibility
- → collaboration
- → scalability
- → security

Disadvantages

- → dependence on internet connection
- → cost
- → performance
- → control

Sample technology and application of web technologies and dbms

- → HTML & CSS
  - HTML – used for creating the primary content of a webpage, giving it structure
  - CSS is the skin that covers html. Its used fore background color, styling, layout, borders, shadowing.
- → Web browsers
  - A web browser takes you anywhere on the internet.
  - It retrieves information from other parts of the web and displays it on your desktop or mobile device.
  - Examples:
    - ➢ Safari – by apple
    - ➢ Google chrome – by google. Cross-platform
    - ➢ Microsoft edge – proprietary. By microsoft
- → Web servers
  - Software and hardware that uses HTTP (hypertext transfer protocol) and other protocols to respond to client requrests made over the www.
  - Main job: to display website content through storing, processing, and delivering web pages to users.
  - Example of client-server model
  - Examples:
    - ➢ Apache – a free open-source web server that delivers web content through the internet
    - ➢ Oracle – a web server designed for medium and large business applications. Was developed originally by Netscape communications corporation in 1996

  - ➢ NGINX – open source web server software used for reverse proxy, load balancing and caching. It provides HTTPS server capabilities and mainly designed for maximum performance and stability.

GROUP 11

Semi structured data and XML

History

- → Pre-1990s
  - Origins of Semistructured Data (Pre-1990s)
  - Before the formalization of semistructured data models, there was a recognition that not all data could be neatly organized into traditional, rigidly structured databases
- → Mid-1990s
  - Development of XML (Mid-1990s)
  - - XML emerged as a response to the need for a flexible, extensible, and standardized way to represent and exchange data over the web
- → Late 1990s
  - XML as a Semistructured Data Format (Late 1990s)
  - XML's hierarchical structure and the ability to define custom schemas made it a natural fit for representing semistructured data.
- → Prime 1990-2000
  - XML in Web Development (Late 1990s - 2000s)
  - XML quickly gained popularity in web development for tasks such as data exchange, configuration files, and representing structured information
- → Early 2000s
  - XML Databases (Early 2000s)
  - The need for databases capable of handling XML data led to the development of XML databases.
  - These databases were designed specifically to store and query XML documents efficiently, recognizing the semistructured nature of the data
- → Prime 2000s
  - XQuery (Early 2000s)
  - XQuery, a query language designed for querying XML data, was developed by the W3C.
  - It provides powerful capabilities for retrieving information from XML documents, making it well-suited for handling the intricacies of semistructured data.
- → 2010
  - NoSQL Movement (2010s)
  - As the demand for handling large volumes of diverse and dynamic data increased, the NoSQL movement gained momentum
- → Prime 2010
  - JSON Competition (2010s)
  - While XML remained prevalent, it faced competition from JSON (JavaScript Object Notation) in certain contexts, especially in web development
- → Ongoing
  - Evolution of XML Standards (Ongoing)
  - XML and related standards, such as XML Schema (XSD), continue to evolve to meet the demands of changing technology landscapes.

Recap:

- → Data is the raw material used to generate meaningful information through processing and analysis
- → Data can be in various forms, including numbers, text, images, audio, video, and more.
- → Data can be quantitative or qualitative, structured or unstructured.

Foundations of Semistructured Data

- → Structured Data – data that has a predefined format, such as a relational database
- → Unstructured Data – data that does not have a predefined format, such as text documents or images
- → Semi-structured data is a type of data that has some level of organization but does not conform to a fixed schema
  - It is often described as "self-describing" because the structure of the data is encoded within the data itself
  - Semi-structured data is often used to represent data that is constantly changing or evolving, such as web pages, social media posts, and sensor readings

Ther semi structured data model (Object exchange Model – OEM)

- → The Object Exchange Model (OEM) is a model for representing semi-structured data.
- → OEM is a nested object model, which means that objects can be nested within other objects.
- → Each OEM object has a unique identifier, a label, a type, and a value.
- → OEM objects can be atomic or complex. Atomic objects contain a single value, while complex objects contain a set of other objects

Characteristics of Semi structured data

- → Self-describing
  - The structure of the data is encoded within the data itself
- → Flexible
  - The structure of the data can change without requiring changes to the schema.
- → Extensible

  - New data elements can be added to the data without requiring changes to the schema.

→ Scalable
  - Semi-structured data can be easily stored and processed in large volumes

Comparison of semi structured data with relational data

→ Semi-structured data has a flexible structure, while relational data has a well-defined structure.

→ Schema can evolve over time, while semi-structured data offers flexibility without affecting the entire dataset.

→ Relational data has a rigid structure and requires careful management.

→ Semi-structured data can be more scalable, while relational data can be challenging to scale horizontally without careful database design

Conclusion

→ Semi-structured data is a growing type of data that is becoming increasingly important in the world of big data.

→ Semi-structured data is more flexible and extensible than structured data, but it can also be more difficult to query and analyze

GROUP 12

Data warehousing Concepts

Business intelligence

→ the collection, methodology, organization, and analysis of data and a software that ingests business data and presents it in user-friendly views such as reports, dashboards, charts and graphs.

Data warehousing

→ a process used to collect and manage data from multiple sources into a centralized repository to drive actionable business insights.

History of BI – Data warehousing concepts

→ Main Idea
  - Data warehouses are designed to support the decision-making process through data collection, consolidation, analytics, and research

→ Early data storage
  - Punch Cards were widely user in 1950s
  - Magnetic Storage in 1960s leading to the popularity of disk storage in 1964

HISTORY

  - The architecture for data warehouses was developed in the 1980s to assist in transforming data from operational systems to decision-making support systems.
    - • Early storage Punch cards – By the 1950s, punch cards were an important pat of the American government and businesses. It is continued to be used regularly until the mid-1980
    - • Magnetic storage - “Magnetic storage” slowly replaced punch cards starting in the 1960s.
    - • Disk storage - Disk storage (hard drives and floppies) started becoming popular in 1964 and allowed data to be accessed directly, significantly improving the clumsier magnetic tapes.

→ Use of NoSQL
  - The emergence of big data led to the development of NoSQL systems, with Facebook adopting a NoSQL system in 2008.
  - NoSQL databases offered scalability advantages in processing big data.

→ Datawarehouse alternatives
  - Data lakes and data lake houses gained popularity for flexible data collection and storage.
  - Data marts were utilized for storing data under the control of specific departments.
  - Data cubes stored data in matrices of three or more dimensions
  - Data silos and swamps represented challenges in large organizations

→ Online applications
  - Commercial online applications emerged, facilitated by direct data access.
  - claims processing, bank teller processing, ATMs, airline reservation processing, retail point-of-sale processing, and manufacturing control processing.

→ PC and 4GL Technology
  - Relational databases gained popularity in the 1980s, offering user□friendly interfaces.
  - Structured Query Language (SQL) became the language for relational database management systems (RDBMS)

→ Need for data warehouses
  - 1990 - with fragmented and inconsistent data due to the of databases and application systems.
  - The internet's popularity, globalization, and increased competition led to the need for true data warehousing.
  - Data warehouses were developed to consolidate data from various databases and support strategic decision-making.

→ Database management systems
  - In 1966, IBM introduced its own DBMS
  - DBMS software capabilities included identifying data locations, resolving conflicts, allowing data

deletion, finding room for data, and quick data retrieval.

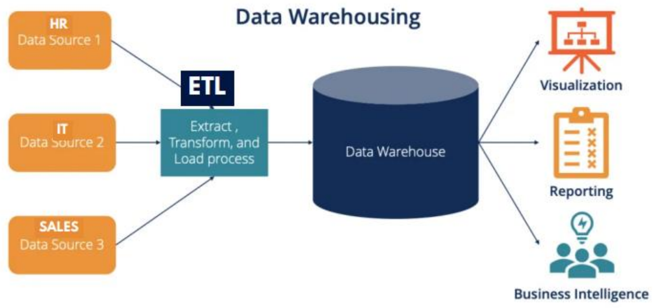
Concept/overview

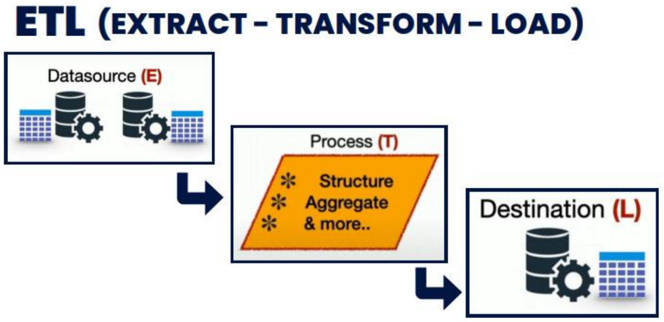
Data warehouse

→ A data warehouse is a type of data management system that is designed to enable and support business intelligence (BI) activities, especially analytics.
→ Data warehouses are solely intended to perform queries and analysis and often contain large amounts of historical data. Because of these capabilities, a data warehouse can be considered an organization's "single source of truth."

Database vs Datawarehouse

→ A database is a transactional system that monitors and updates real-time data in order to have only the most recent data available
→ A data warehouse is programmed to aggregate structured data over time.





Features of data warehouse

→ Subject-oriented
  - Data is organized around specific business subjects
→ Integrated
  - Data from multiple sources is integrated and reconciled to ensure consistency and accuracy
→ Time-variant
  - Data stores historical information over time
→ Nonvolatile
  - All data is read-only and not subject to changes from operational systems.

Benefits of data warehouse

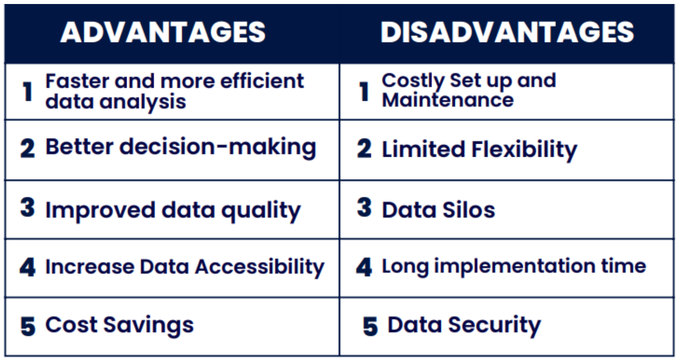
→ Improve business intelligence and efficiency
→ Save time and enhance decision making speed
→ Increase data security
→ Maintain historical data for long-term insights

Advantages of data warehouse

→ Faster and more efficient data analysis
→ Better decision-making
→ Improved data quality
→ Increased data accessibility
→ Cost savings

Disadvantages of data warehouse

→ Costly setup and maintenance
→ Limited flexibility
→ Data silos
→ Long implementation time
→ Data security



ADVANTAGES	DISADVANTAGES
1 Faster and more efficient data analysis	1 Costly Set up and Maintenance
2 Better decision-making	2 Limited Flexibility
3 Improved data quality	3 Data Silos
4 Increase Data Accessibility	4 Long implementation time
5 Cost Savings	5 Data Security

Data warehousing tools

→ Hevo data
→ Oracle autonomous data warehouse
→ Google data warehouse tools
→ Amazon redshift
→ Microsoft azure data warehouse tools
→ Snowflake

Applications of data warehousing

→ Manufacturing & supply chain
  - Data warehouses can help in inventory management, all the data related to vendors, logistics , and ultimately serving the customer better.
→ Banking & finance
  - Data security is critical for the BFSI sector, and data warehouses solve that problem by vouching for industry-standard security compliances. The warehouses can be used to get updates about customer deposits, loans, funds, deposits, etc., and a better understanding of the performance of different branches

→ E-commerce
  - E-commerce platforms need to gather key marketing metrics (such as clicks, impressions, website visitors, etc.) from marketing tools and use that to approach their customers in a better way.
  - This is where data warehouses help. Replicating data, tracking & visualizing KPIs such as conversion rates, churn rates, and return on ad spends, safe storage, etc. help companies perform better.

→ Healthcare
  - Data warehouses are always working to digitally improve medical infrastructure, minimize wait time, and simplify processes.
  - Getting personalized healthcare can be possible with a single platform (such as having one place for all diagnostics, tests, prescriptions, and follow-ups).
  - The warehouse stores all clinical, financial, and employee data, which is analyzed to maximize resource allocation.

→ Financial auditing
  - With access to real-time financial data, warehouses ensure decisions related to the business's current financial performance can be reached quickly.

→ Pharmaceuticals
  - As data warehouses make data more accessible, it's now being used for making better strategic decisions and identifying & developing customer buying trends in pharmaceuticals

GROUP 13

Data warehouse design

→ DWH design is about creating a relational database schema for storing and analyzing large volumes of data from various sources.

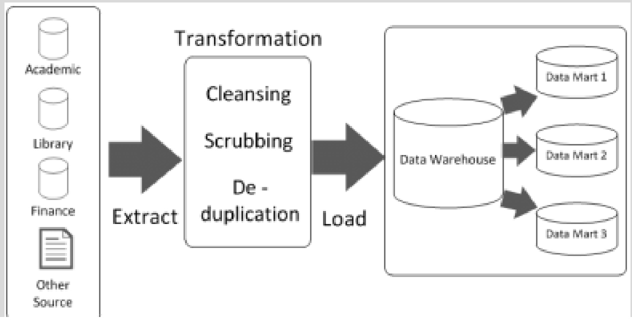
→ Data warehouse design takes a method different from view materialization in the industries.

→ It sees data warehouses as database systems with particular needs such as answering management related queries.

→ The purpose of a data warehouse is to provide a centralized repository of data that can be accessed and analyzed by business analysts, data scientists, and other stakeholders.

→ Data warehouse design is essential because it provides a structured and organized way of storing and analyzing large amounts of data from various sources. Here are some reasons that suggest the importance:
  - Improved data quality
  - Better decision making
  - Efficient Querying
  - Integration of data
  - Scalability

Top-down approach



→ In the top-down design approach, data warehouse is described as a subject-oriented, time-variant, non-volatile and integrated data repository for the entire enterprise data from different sources are validated, reformatted, and saved in a normalized database as the data warehouse

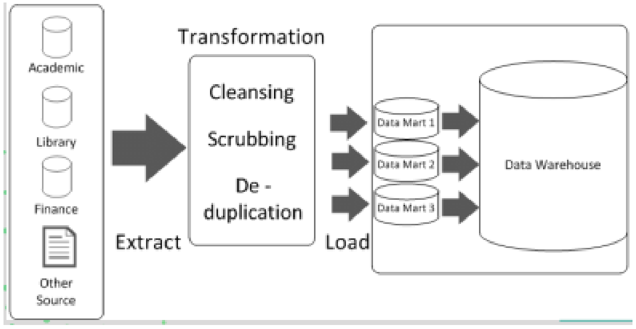
→ The data warehouse stores "atomic" information, the data at the lowest level of granularity, from where dimensional data marts can be built by selecting the data required for specific business subjects or particular departments

Advantage	Disadvantage
This approach is robust against business changes. Creating a new data mart from the data warehouse is very easy	expensive
Its easier to maintain top-down design	Time-consuming
Provides consistent dimensional views of data across data marts, as all data marts are loaded from data warehouse	Highly skilled people required for set up

What companies use the top-down approach

→ Amazon
→ J.P. Morgan
→ IBM
→ Google

Bottom-up approach



- → A bottom-up approach to data designing, also known as data-driven or reverse engineering
- → Starts with the most detailed and specific level of data such as tables, columns, and attributes and then builds up to higher levels of abstraction
- → Can be useful when you need to understand the current state of data in a system or organization., identify data quality issues and inconsistencies, optimize data storage and performance, or implement data governance and security policies

advantages	Disadvantages
Has consistent data marts and these data marts can be delivered quickly	It is difficult to maintain and often redundant and subject to revisions
It takes less time	This model is not strong as top down approach as dimensional view of data marts is not consistent
Data marts are created first to provide reporting capability	The bottom up approach can be slow, as each component must be completed before moving on to the next

What companies use the bottom-up approach

- → Hp
- → Toyota
- → Oracle
- → Samsung

Group 14

OLAP and data mining

OLAP

- → Online Analytical Processing (OLAP)
- → s a type of decision support system (DSS) that allows users to analyze data from multiple dimensions. It is designed to provide fast, interactive access to large amounts of data, allowing users to drill down, roll up, and slice data to gain insights

Data mining

- → Data mining is the process of extracting knowledge from data.
- → It is a broad field that encompasses a variety of techniques, such as classification, clustering, and association rule mining.
- → Data mining has its roots in statistics and machine learning.

History of OLAP

- → Early 1970s
  - Edgar Codd, the inventor of the relational model, proposed a new data model called the Multidimensional Model.
- → 1980s
  - Several commercial OLAP products were developed, including Essbase and Express.
- → 1990s
  - OLAP standards were developed, such as OLE DB for OLAP and XML for analysis.
- → Present
  - OLAP is a mature technology that is used by businesses of all sizes to analyze data.

History of data mining

- → 1960s & 1970s
  - Researchers developed algorithms for classification and clustering, which are two of the most fundamental techniques in data mining.
- → 1980s
  - Data mining became a more recognized field, and researchers began to develop new techniques for extracting knowledge from data
- → 1990s
  - Data mining became a commercial reality, and several data mining tools were developed.
- → Present
  - Data mining is a widely used technology that is used by businesses of all sizes to gain insights from their data

Features of OLAP

- → Multidimensional analysis
- → Fast data retrieval
- → Visualization and reporting
- → Aggregation and summarization

Features of data mining

- → Pattern recognition
- → Knowledge discovery
- → Predictive modeling
- → Automation and scalability

Types of OLAP and data mining

→ OLAP
  - Supervised learning
  - Unsupervised learning
→ Data mining
  - Real-time OLAP (ROLAP)
  - Multidimensional OLAP (MOLAP)
  - Hybrid OLAP (HOLAP)

Relationship between OLAP and data mining

→ OLAP and data mining are complementary techniques that serve different purposes in the data analysis process
→ OLAP provides a platform for analyzing data from multiple dimensions
→ while data mining provides a set of techniques for extracting knowledge from data.
→ OLAP can be used to prepare data for data mining.
→ Data mining can be used to analyze data that has been prepared with OLAP

Benefits of OLAP

→ Faster decision making
→ Non-technical user support
→ Integrated data view
→ Self-service reporting

Benefits of data mining

→ New insights
→ Predictions
→ Decision support systems

Advantages of OLAP

→ Fast query performance
→ Multidimensional analysis
→ Ease of use

Disadvantages of OLAP

→ Complexity
→ Cost

Advantages of data mining

→ Ability to discover new insights
→ Ability to make predictions
→ Ability to develop decision support systems

Disadvantages of data mining

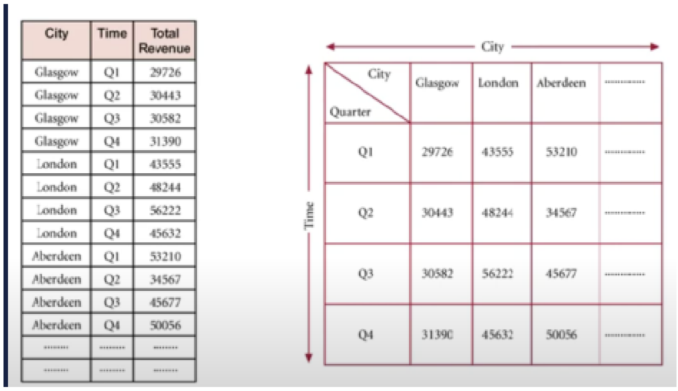
→ Complexity
→ Cost
→ Potential for bias

OLAP / Data mining applications

→ Business reporting for sales
  - The Business Reporting gives an overview of the sales activity in the sales activities within an organization. It shows the trends in the sales over a certain time period.
→ Marketing
  - Industries like digital marketing, health care, eCommerce, and finance uses OLAP in their marketing.
→ Management reporting
  - It aims to inform the managers of different aspects of the organizations about the data from the various departments of the company in order to help them to make better decisions.
→ Financial reporting
  - Financial Reporting refers to financial reports of an organization that are released to stakeholders and the public. It includes the financial statements which include the balance sheet, income sheet, statement of cash flows, etc

OLAP Applications

→ Multi-dimensional views of data – 2 dimensional view



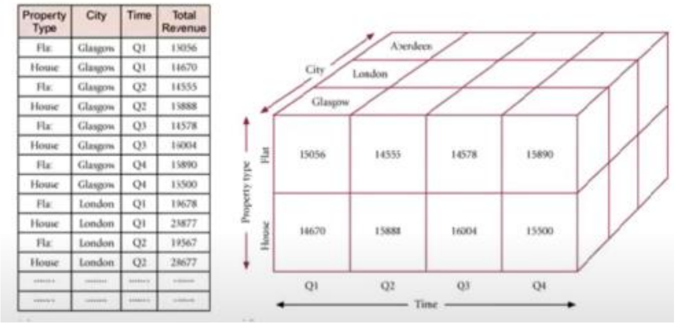
City	Time	Total Revenue
Glasgow	Q1	29726
Glasgow	Q2	30443
Glasgow	Q3	30582
Glasgow	Q4	31390
London	Q1	43555
London	Q2	48244
London	Q3	56222
London	Q4	45632
Aberdeen	Q1	53210
Aberdeen	Q2	34567
Aberdeen	Q3	45677
Aberdeen	Q4	50056
....	....	....
....	....	....

City

Time

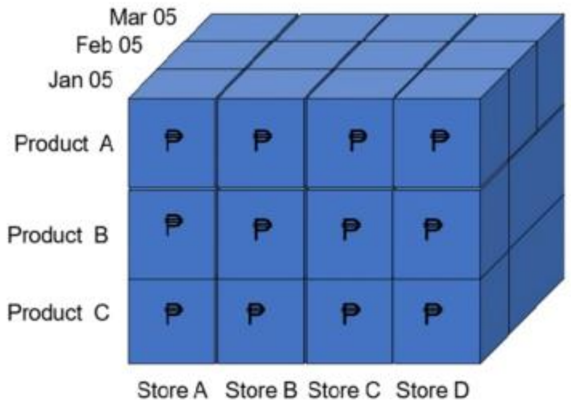
Quarter \ City	Glasgow	London	Aberdeen	....
Q1	29726	43555	53210	....
Q2	30443	48244	34567	....
Q3	30582	56222	45677	....
Q4	31390	45632	50056	....

→ Multi-dimensional views of data - 3 dimensional view



Property Type	City	Time	Total Revenue
Flat	Glasgow	Q1	15056
House	Glasgow	Q1	14670
Flat	Glasgow	Q2	14555
House	Glasgow	Q2	15888
Flat	Glasgow	Q3	14578
House	Glasgow	Q3	16004
Flat	Glasgow	Q4	15890
House	Glasgow	Q4	15500
Flat	London	Q1	19678
House	London	Q1	23877
Flat	London	Q2	19567
House	London	Q2	28677
....	....	....	....
....	....	....	....

→ Data cube



OLAP Operations

→ Slice
- Select data on a single dimension of a data cube

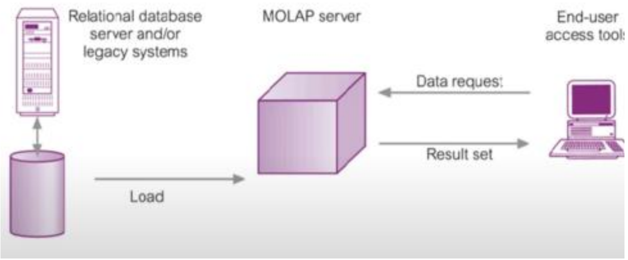
→ Dice
- Extracts a sub-cube from the original cube

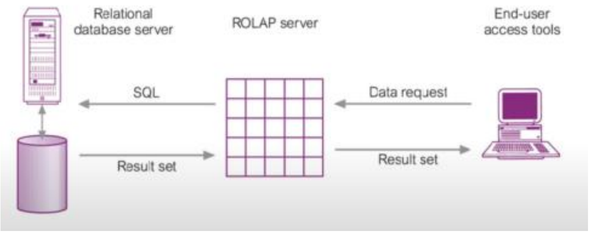
→ Roll-up
- Combining of cells for one dimension

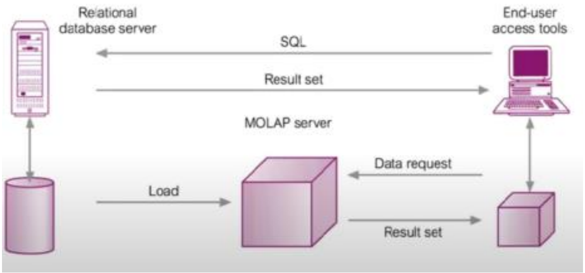
→ Drill-down
- Reverse of “Roll-up” operation

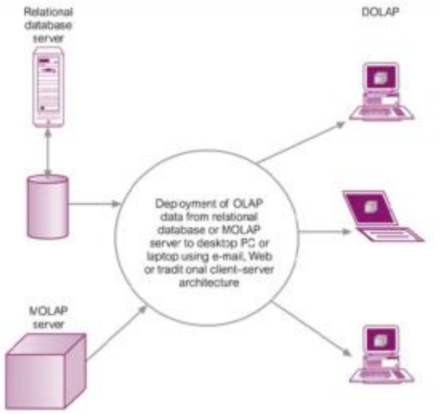
→ Rotation
- Allow user to view data from a new perspective

OLAP tools

→ Multi-dimensional OLAP (MOLAP)
- MOLAP Architecture
- 

→ Relational OLAP (ROLAP)
- 

→ Hybrid OLAP (HOLAP)
- 

→ Desktop OLAP (DOLAP)
- 
Relational database server
DOLAP
Deployment of OLAP data from relational database or MOLAP server to desktop PC or laptop using e-mail, Web or traditional client-server architecture
MOLAP server

Data Mining Tools

→ WEKA

→ ORANGE