SUBJECT CODE : 410253(C)

As per Revised Syllabus of

SAVITRIBAI PHULE PUNE UNIVERSITY

Choice Based Credit System (CBCS)

B.E. (Computer) Semester - VIII (Elective VI)

BUSINESS INTELLIGENCE

Iresh A. Dhotre

M.E. (Information Technology)

Ex-Faculty, Sinhgad College of Engineering

Pune

Dr. Sarika N. Zaware

Ph.D (Computer Science and Engg.)

M.E. (Computer Science and Engg.)

MBA (Computer specialization)

B.E. (Computer Engineering)

Professor, Department of Computer Engineering

AISSMS Institute of Information Technology, Pune

TECHNICAL PUBLICATIONS

An Up-Thrust for Knowledge

BUSINESS INTELLIGENCE

Subject Code : 410253(C)

B.E. (Computer Engineering) Semester - VIII (Elective VI)

Published by :

TECHNICAL PUBLICATIONS

Printed :

SYLLABUS

Business Intelligence - (410253(C))

Credit	Examination Scheme :
03	In-Sem (Paper) : 30 Marks End-Sem (Paper) : 70 Marks

Unit I Introduction to Decision support systems and Business intelligence

Decision support systems : Definition of system, representation of the decision-making process, evolution of information systems, Decision Support System, Development of a decision support system, the four stages of Simon's decision-making process and common strategies and approaches of decision makers

Business Intelligence : BI, its components and architecture, previewing the future of BI, crafting a better experience for all business users, End user assumptions, setting up data for BI, data, information and knowledge. The role of mathematical models. Business intelligence architectures. Ethics and business intelligence **(Chapter - 1)**

Unit II The Architecture of DW and BI

BI and DW architectures and its types - Relation between BI and DW - OLAP (Online analytical processing) definitions - Different OLAP Architectures - Data Models-Tools in Business Intelligence - Role of DSS, EIS, MIS and digital Dash boards - Need for Business Intelligence

Difference between OLAP and OLTP - Dimensional analysis - What are cubes ? Drill-down and roll-up - slice and dice or rotation - OLAP models - ROLAP versus MOLAP - defining schemas : Stars, snowflakes and fact constellations. **(Chapter - 2)**

Unit III Reporting Authoring

Building reports with relational vs Multidimensional data models, Types of Reports - List, crosstabs, Statistics, Chart may, financial etc. Data Grouping and Sorting, Filtering Reports, Adding Calculations to Reports, Conditional formatting, Adding Summary Lines to Reports Drill up, drill-down, drill-through capabilities, Run or schedule report, different output forms - PDF, excel, csv, xml etc **(Chapter - 3)**

Unit IV Data preparation

Data validation : Incomplete data, Data affected by noise **Data transformation** : Standardization, Feature extraction. **Data reduction** : Sampling, Feature selection, Principal component analysis, Data discretization. **Data exploration** : **1. Univariate analysis** : Graphical analysis of categorical attributes

Graphical analysis of numerical attributes, Measures of central tendency for numerical attributes, Measures of dispersion for numerical attributes, Identification of outliers for numerical attributes **2. Bivariate analysis** : Graphical analysis, Measures of correlation for numerical attributes, Contingency tables for categorical attributes, **3. Multivariate analysis** : Graphical analysis, Measures of correlation for numerical attributes **(Chapter - 4)**

Unit V Impact of Machine learning in Business Intelligence Process

Classification : Classification problems, Evaluation of classification models, Bayesian methods, Logistic regression, **Clustering** : Clustering methods, Partition methods, Hierarchical methods, Evaluation of clustering models, **Association Rule** : Structure of Association Rule, Apriori Algorithm **(Chapter - 5)**

Unit VI BI Applications

Tools for Business Intelligence, Role of analytical tools in BI, Case study of Analytical Tools WEKA, KNIME, Rapid Miner, R: Data analytics, Business analytics, ERP and Business Intelligence, BI and operation management, BI in inventory management system, BI and human resource management, BI Applications in CRM, BI Applications in Marketing, BI Applications in Logistics and Production, Role of BI in Finance, BI Applications in Banking, BI Applications in Telecommunications, BI in sales force management **(Chapter - 6)**

Unit I

1 Introduction to Decision Support Systems and Business Intelligence

Syllabus

Decision support systems : *Definition of system, representation of the decision-making process, evolution of information systems, Decision Support System, Development of a decision support system, the four stages of Simon's decision-making process, and common strategies and approaches of decision makers*

Business Intelligence : *BI, its components & architecture, previewing the future of BI, crafting a better experience for all business users, End user assumptions, setting up data for BI, data, information and knowledge, The role of mathematical models, Business intelligence architectures, Ethics and business intelligence*

Contents

1.1 Definition of System

- Term *system* is often used in everyday language. For example, we refer to the solar system, the nervous system or the justice system. System share some common characteristics.
- System can be open or closed. For open system, its boundaries can be crossed in both directions by flows of materials and information. When such flows are lacking, the system is said to be closed.
- Fig. 1.1.1 shows abstract representation of a system.

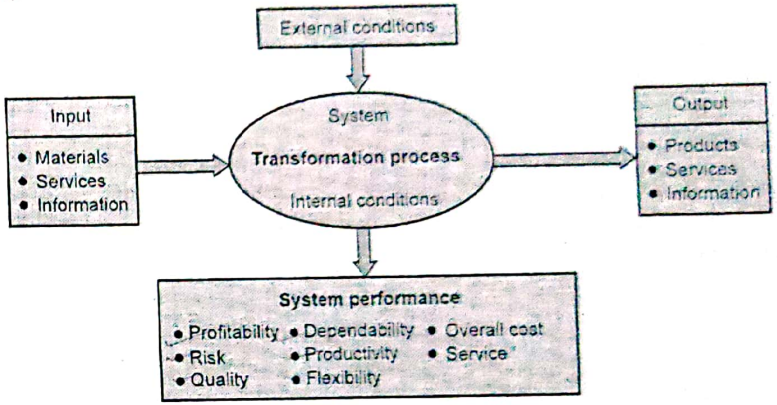


Fig. 1.1.1 Abstract representation of a system

- Input to system is number of parameters like service, material and information. It is transform to output using transformation process regulated by internal conditions and external conditions.
- A system will often incorporate a feedback mechanism. Feedback occurs when a system component generates an output flow that is fed back into the system itself as an input flow, possibly as a result of a further transformation.
- Fig. 1.1.2 shows closed cycle system. Systems that are able to modify their own output flows based on feedback are called closed cycle systems.

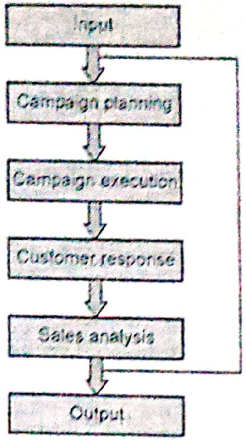


Fig. 1.1.2 Closed cycle marketing system with feedback effects

- Evaluation metrics are of two types : Effectiveness and efficiency.
  1. Effectiveness : Effectiveness measurements express the level of conformity of a given system to the objectives for which it was designed. Effectiveness mean doing the right action.
  2. Efficiency : Efficiency means doing the action in best possible way. It measurements highlight the relationship between input flows used by the system and the corresponding output flows. Efficiency measurements are therefore associated with the quality of the transformation process.

1.2 Representation of the Decision-Making Process

- Decision making refers to making choices among alternative courses of action, which may also include inaction. A decision is a choice from multiple alternatives, usually made with a fair degree of rationality.
- Decision making is the process of making choices by identifying a decision, gathering information and assessing alternative resolutions. Effective and successful decisions make profit to the company and unsuccessful ones make losses. Therefore, corporate decision making process is the most critical process in any organization.
- In order to build effective decision support systems, we first need to describe in general terms how a decision making process is articulated.

1. Rationality and problem solving
2. The decision making process
3. Types of decisions
4. Approaches to the decision making process

1.2.1 Rationality and Problem Solving

- The rational decision-making model describes a series of steps that decision makers should consider if their goal is to maximize the quality of their outcomes. Decision-making and problem-solving are inter-related. It is the process through which managers identify organizational problems and solve them.
- Decisions may be major or minor, strategic or operational, long-term or short-term. They are made for each functional area at each level.
- Factors influencing a rational choice :
  1. Economic : Most influential factor in decision making processes. Aim is minimization of costs or the maximization of profits.
  2. Technical : Options that are not technically feasible must be discarded. For example, production plan that exceeds the maximum capacity of a plant cannot be regarded as a feasible option.
  3. Legal : Legal rationality implies that before adopting any choice the decision makers should verify whether it is compatible with the legislation in force within the application domain
  4. Ethical : Decision should abide by the ethical principles and social rules of the community to which the system belongs.
  5. Procedural : A decision may be considered ideal from an economic, legal and social standpoint, but it may be unworkable due to cultural limitations of the organization.

1.2.2 The Decision-Making Process

- Characterizing a decision-making process :
  1. Decisions are often devised by a group of individuals instead of a single decision maker.
  2. The number of alternative actions may be very high and sometimes unlimited.
  3. The effects of a given decision usually appear later, not immediately.

4. The decisions made within a public or private enterprise or organization are often interconnected and determine broad effects. Each decision has consequences for many individuals and several parts of the organization.
5. During the decision-making process knowledge workers are asked to access data and information and work on them based on a conceptual and analytical framework.
6. Feedback plays an important role in providing information and knowledge for future decision-making processes within a given organization.
7. In most instances, the decision-making process has multiple goals, with different performance indicators, that might also be in conflict with one another.
8. Many decisions are made in a fuzzy context and entail risk factors.

1.2.3 Types of Decisions

- Decisions are divided into two types : nature and scope.

1. Type of decision by nature :

- Structured, unstructured or semi-structured are types of decision as per nature.
- **Structured decisions :** A decision is structured if it is based on a well-defined and recurring decision-making procedure. Structured decisions, by contrast, are repetitive and routine, and decision makers can follow a definite procedure for handling them to be efficient.
- **Unstructured decisions :** A decision is said to be unstructured if the three phases of intelligence, design and choice are also unstructured. Unstructured decisions are those in which the decision maker must provide judgment, evaluation and insights into the problem definition
- **Semi-structured decisions :** A decision is semi-structured when some phases are structured and others are not. Semi-structured decisions are those in which only part of the problem has a clear-cut answer provided by an accepted procedure. In general, structured decisions are more prevalent at lower organizational levels and unstructured decision making is more common at higher levels.
- Senior executives face many unstructured decision situations, such as establishing the firm's five or ten-year goals. Middle management faces more structured decision scenarios but their decisions may include unstructured components.

Operational management and rank-and-file employees tend to make more structured decisions.

2. Types of decision by scope

- Depending on their scope, decisions can be classified as *strategic*, *tactical* and *operational*.
- Fig. 1.2.1 shows support for various decision-making levels

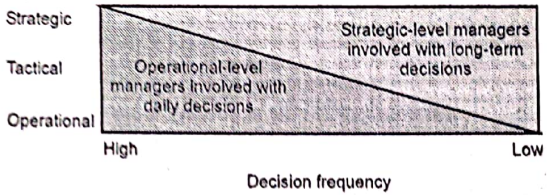


Fig. 1.2.1 Support for various decision - making levels

a) Strategic Decisions and Plans

- A board of directors, whose members are elected by a company's shareholders, makes strategic decisions for a company. Strategic decisions are decisions and plans that have long-term or material impact on a company. These often include, but are not limited to, decisions regarding :
  i. The election of officers and executives
  ii. Equity grants or transfers (including compensation packages for officers and executives)
  iii. Annual budgets and audits
  iv. Amendments to the certificate of incorporation or bylaws
  v. Shareholder distributions
  vi. Sale of substantially all a company's assets
  vii. Dissolution or sale of a company
  viii. Adoption of employee benefit plans
  ix. Other materially important agreements or long-term strategies

b) Tactical Decisions and Plans

- Officers and executives make tactical decisions for a company. Officers and executives include the CEO, COO, CFO and other top-level management in a company.

- Tactical decisions are decisions and plans that concern the more detailed implementation of the directors' general strategy, usually with a medium-term impact on a company.
- Tactical points requiring decisions include, but are not limited to :
  i. Size and structure of a work force
  ii. Sales and marketing strategy
  iii. Signing non-disclosure agreements
  iv. Work assignments allocated to particular groups and people
  v. Large purchases within a previously approved budget

c) Operational Decisions

- Operational managers and other employees make operational decisions for the company. Operational managers include mid-level, supervisory and lower-level management.
- Operational decisions are the day-to-day decisions that have only a short-term impact on a company. These include, but are not limited to, decisions regarding :
  i. Scheduling employees
  ii. Training on specific tasks in a company (e.g., sales techniques, computer training, etc.)
  iii. Purchasing office supplies
  iv. Assigning work to specific employees

1.3 Evolution of Information Systems

- An information system is a set of computerized components that are used to collect, create and store data. It is also used to process the data into information and distribute it to the desired destinations.
- The first business application of computers (in the mid- 1950s) performed repetitive, high-volume, transaction-computing tasks. The computers" crunched numbers" summarizing and organizing transactions and data in the accounting, finance and human resources areas. Such systems are generally called Transaction Processing Systems (TPSs).
- Management Information Systems (MISs) : These systems access, organize, summarize and display information for supporting routine decision making in the functional areas.

- Office Automation Systems (OASs) : Such as word processing systems were developed to support office and clerical workers.
- Decision Support Systems : Were developed to provide computer based support for complex, non-routine decision.
- End- user computing : The use or development of information systems by the principal users of the systems' outputs, such as analysts, managers, and other professionals.
- Intelligent Support System (ISSs) : Include expert systems which provide the stored knowledge of experts to non-experts and a new type of intelligent system with machine- learning capabilities that can learn from historical cases.
- Knowledge management systems : Support the creating, gathering, organizing, integrating and disseminating of organizational knowledge.
- Data warehousing : A data warehouse is a database designed to support DSS, ESS and other analytical and end-user activities.
- Mobile computing : Information systems that support employees who are working with customers or business partners outside the physical boundaries of their company; can be done over wire or wireless networks.

1.4 Decision Support System

- A decision support system (DSS) is an interactive computer-based application that combines data and mathematical models to help decision makers solve complex problems faced in managing the public and private enterprises and organizations.
- A decision support system may present information graphically and may include an expert system or Artificial Intelligence (AI). It may be aimed at business executives or some other group of knowledge workers.
- DSS is an interactive computer system helping decision makers to combine data and models to solve semi-structured and unstructured problems. Fig. 1.4.1 shows structure of a decision support system.

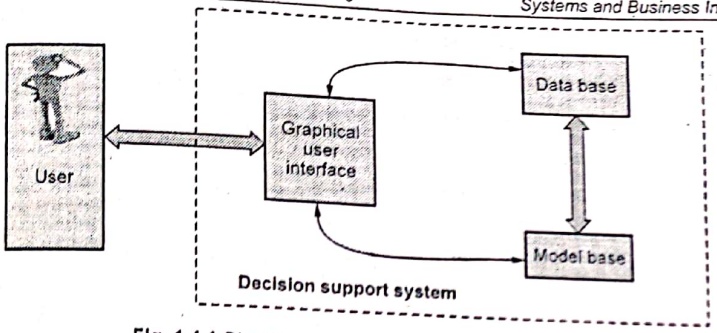


Fig. 1.4.1 Structure of a decision support system

- A good DSS helps decision-makers with compiling various types of data gathered from several different sources, including documents, raw data, management, business models and personal knowledge from employees.
- Decision makers perform quantitative analysis on data. Model base contains a list of models for mathematical computations.
- It is a component of a business intelligence system and usually includes a database of information related to the specific industry. Decision support systems can also include files, company models, health information, sales data, projections, marketing numbers and personal knowledge. Industries like healthcare, agriculture and marketing frequently take advantage of decision support systems.
- Every decision support system has a database. The database is usually the primary aspect of the decision support system that makes it useful for an organization, as the program can examine large stores of data when supporting decisions much faster than a single person or team could. The database's information depends on the category of system and industry the system is for. Some databases may contain statistics, while others may be more document-focused.
- Models : Decision support systems also create models to support professionals in taking action that positively affects their situation. The models within the decision support system are the predictions or trajectories the program determines are plausible
- DSS is a generic concept that describes information systems that provide analytical modelling and information to support semi-structured and unstructured organizational decision making.

- DSS software system : It consists of various mathematical and analytical models that are used to analyze the complex data, thereby producing the required information. A model predicts the outputs in the basis of different input or different conditions, or finds out the combination of conditions and input that is required to produce the desired output. A decision support system may compromise different models where each model performs a specific function.
- User interface : It is an interactive graphical interface which makes the interaction easier between the DSS and its users. It displays the results of the analysis in various forms, such as text table, charts or graphics. The user can select appropriate option to view output according to his requirement.
- Fig. 1.4.2 shows extended DSS.

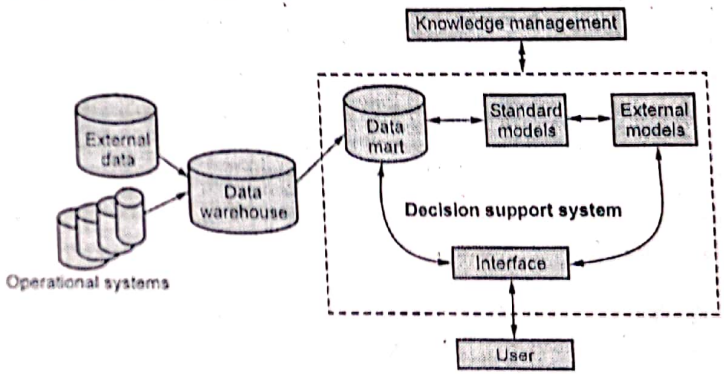


Fig. 1.4.2 Extended DSS

- Data warehouse : Database generally provides current information about the organization relating to the underlying transactional processes, but it fails to provide historical, contend rich information that are often more important to the decision making process than standalone islands of information.
- The data warehouse fills this gap by capturing operational data and presenting it in a more meaningful format, using a relational database and ultimately complimenting the functions of the DB used in the DSS. Thus the data warehouse and the DB coexists to provide synergistic outcomes which supports information requirement of the DSS superimposed on the systems platform.
- Data mart is a subset of the data warehouse. Data mart is usually assigned to a specific business unit within the enterprise. Data mart is used to slice data

warehouse into a different business unit. Typically, ownership of the data mart is given to that particular business unit or department.

- The Decision Support Systems can be divided into following categories :

1. **Model-driven DSS :** A model-driven DSS was based on simple quantitative models. It used limited data and emphasized manipulation of financial models. A model-drive DSS was used in production planning, scheduling and management. It provided the most elementary functionality to manufacturing concerns.
2. **Data-driven DSS :** Data-driven DSS emphasized the access and manipulation of data tailored to specific tasks using general tools. While it also provided elementary functionality to businesses, it relied heavily on time-series data. It was able to support decision making in a range of situations.
3. **Communication-driven DSS :** As the name suggests, communication-driven DSS uses communication and network technologies to facilitate decision making. The major difference between this and the previous classes of DSS was that it supported collaboration and communication. It made use of a variety of tools including computer-based bulletin boards, audio and video conferencing.
4. **Document-driven DSS :** A document-driven DSS uses large document databases that stores documents, images, sounds, videos and hypertext docs. It has a primary search engine tool associated for searching the data when required. The information stored can be facts and figures, historical data, minutes of meetings, catalogs, business correspondences, product specifications, etc.
5. **Knowledge-driven DSS :** Knowledge-based DSS are human-computer systems that come with a problem-solving expertise. These combine artificial intelligence with human cognitive capacities and can suggest actions to users. The notable point is that these systems have expertise in a particular domain.
6. **Web-based DSS :** Web-based DSS is considered most sophisticated decision support system that extends its capabilities by making use of worldwide web and internet. The evolution continues with advancement in internet technology.

1.4.1 Characteristics and Capabilities of DSS

- Common characteristics of DSS :
  1. Uses underlying data and model
  2. Handles large amounts of data from different sources
  3. Provides report and presentation flexibility
  4. Offers both textual and graphical orientation
  5. Problem structure, used in semi-structured and unstructured decision context.
  6. Uses underlying data and model
- **Capabilities of DSS**
  1. Provide support in semi-structured and unstructured situations, includes human judgment and computerized information
  2. Support for various managerial levels
  3. Support to individuals and groups
  4. Support to interdependent and/or sequential decisions
  5. Support all phases of the decision-making process
  6. Support a variety of decision-making processes and styles
  7. Have user friendly interfaces
  8. Goal : Improve effectiveness of decision making
  9. The decision maker controls the decision-making process
  10. End-users can build simple systems
  11. Utilizes models for analysis
  12. Provides access to a variety of data sources, formats and types

1.4.2 Advantages and Disadvantages of DSS

Advantages :

- Increase organizational control.
- Increase decision maker satisfaction.
- Improve interpersonal communication. DSS can improve communication and collaboration among decision makers.
- Increasing productivity

Disadvantages :

- Information overload : A computerized decision making system may sometimes result in information overload.
- Overemphasis on decision making
- Cost of development

1.4.3 Application of DSS

1. Agriculture : Farmers use DSS tools for crop planning to help determine the best planting time, as well as when to fertilize and when to harvest.
2. Medicine : When a DSS is used in medicine, it is known as a clinical DSS. The technology can be used in a variety of ways, such as maintaining research information about chemotherapy protocols, preventive and follow-up care and monitoring medication orders. DSSs are also used with medical diagnosis software.
3. Weather forecasting : Some states have used DSSs to provide information about potential hazards, such as floods. The system includes real-time weather conditions and may include information, both current and historical, about floodplain boundaries and county flood data.
4. Real estate : Real estate companies often use DSSs to manage data on comparable home prices and acreage.
5. Education : Universities and colleges rely on DSSs to know how many students are currently enrolled. This helps them predict how many students will register for particular courses or whether the student population is sufficient to meet the university's costs.

1.5 Development of a Decision Support System

- Fig. 1.5.1 shows the major steps in the development of a DSS.
- Planning : This phase understand user requirement and opportunities. Planning usually involves a feasibility study to address the question : Why do we wish to develop a DSS? During the feasibility analysis, general and specific objectives of the system, recipients, possible benefits, execution times and costs are laid down.
- Analysis : In the analysis phase, it is necessary to define in detail the functions of the DSS to be developed, by further developing and elaborating the preliminary conclusions achieved during the feasibility study.
- Design : The entire architecture of the system is defined, through the identification of the hardware technology platforms, the network structure, the software tools to develop the applications and the specific database to be used.

- Implementation : It relates to the overall impact on the organization determined by the new system.

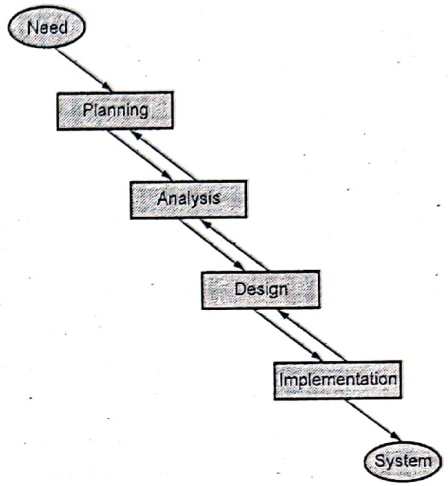


Fig. 1.5.1 Phases of DSS

1.6 The Four Stages of Simon's Decision-Making Process

- Simon's model defines four phases of decision-making process : Intelligence Phase, Design Phase, Choice Phase and Implementation Phase.
- Fig. 1.6.1 shows four stages of Simon's decision-making process

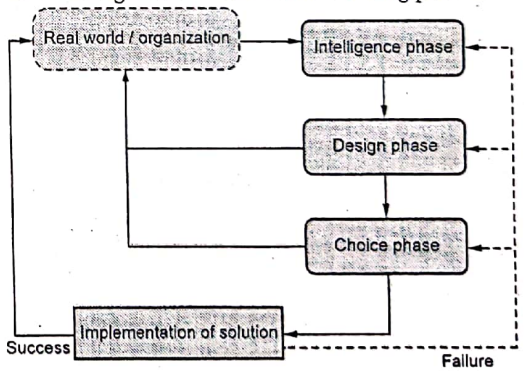


Fig. 1.6.1 Four stages of Simon's decision-making process

- There are four different stages in decision making :
  1) Intelligence : Consists of identifying and understanding a problem. Business Intelligence implementations are considered successful only if you have clear business needs and see real benefits from it. Business Intelligence is not just about data. It should be connected with organizational goals and objectives.
- Scan the environment, either intermittently or continuously
- Identify problem situations or opportunities
- Monitor the results of the implementation
- Problem is the difference between what people desire (or expect) and what is actually occurring
- Timely identification of opportunities is as important as identification of problems
  2) Design : Involves exploring various solutions. The main goal of the design phase is to define and construct a model which represent a system, by defining relationships between collected variables.
- Finding/developing and analyzing possible courses of actions
- A model of the decision-making problem is constructed, tested and validated
- Modeling : Conceptualizing a problem and abstracting it into a quantitative and/or qualitative form (i.e., using symbols/variables)
  a. Abstraction : Making assumptions for simplification
  b. Tradeoff (cost/benefit) : More or less abstraction
  c. Modeling : Both an art and a science
  3) Choice : Consists of choosing among available solutions. In this phase we are actually making decisions. The end product of this phase is a decision. Decision is made by selecting and evaluating alternatives defined in previous step.
- The actual decision and the commitment to follow a certain course of action are made here
- The boundary between the design and choice is often unclear (partially overlapping phases)
- Includes the search, evaluation, and recommendation of an appropriate solution to the model
- Solving the model versus solving the problem
  4) Implementation Phase : All the previous steps we have made (intelligence, design and choice) are now implemented. Implementation can be either

successful or not. Successful implementation results with a solution to the defined problem. On the other hand, failure brings us back to the earlier phase.

1.7 Common Strategies and Approaches of Decision Makers

- Leadership responsibility is to take great decisions. If we select the wrong decision making style, we could face a disaster. So choose the right style and make decision faster and more effectively.
  1. Autocratic Decision-Making : For situations where we have low impact and they're reasonably small decisions, but they get larger as urgency goes up, an Autocratic decision-making style is the most appropriate. In Autocratic decision-making, decisions are made at the top. Actually, it may be counterproductive to involve a lot of people in making the call.
- Typically, in an environment where we are making Autocratic decisions, work activities and roles are very tightly structured, they're monitored and well controlled. Command and control is very important in these situations.
  2. Participatory Decision-Making : For larger decisions where there's higher urgency and we need to make a call soon, but the impact is going to be big, we are looking at a situation where we need to use a Participatory decision-making style. This is where we are going to make a decision with input from the people who are going to be impacted in that final call. Participatory decisions are made when the decision is much bigger and there's a lot more risk involved.
  3. Consensus-Based Decision-Making : For situations where it's a large decision but there's no urgency around it and we have got plenty of time, we can be using a Consensus-based decision-making style. This is where decisions are reached with a cross-functional team.
- People from different departments have input, and buy-in is essential.
  4. Democratic Decision-Making : For mid-sized decisions where there's not a lot of urgency but we do need to make a decision and move on, a Democratic style is the most appropriate. This is where a decision is reached by a majority vote. Buy-in is desirable but it's not essential.

1.8 Business Intelligence

- Business Intelligence (BI) is a business management term which refers to applications and technologies which are used to gather, provide access to and analyze data and information about company operations.

- Business intelligence systems can help companies have a more comprehensive knowledge of the factors affecting their business, such as metrics on sales, production, internal operations and they can help companies to make better business decisions.
- Business intelligence may be defined as a set of mathematical models and analysis methodologies that exploit the available data to generate information and knowledge useful for complex decision-making processes.
- Business intelligence combines business analytics, data mining, data visualization, data tools and infrastructure and best practices to help organizations make more data-driven decisions.
- Business intelligence includes data analytics and business analytics but uses them only as parts of the whole process. BI helps users draw conclusions from data analysis. Data scientists dig into the specifics of data, using advanced statistics and predictive analytics to discover patterns and forecast future patterns.

1.8.1 Components of BI

- Fig. 1.8.1 shows components of a business intelligence system. The main components of a business intelligence system decisions, optimization, data mining, data exploration, Data warehouse/Data mart and data source.

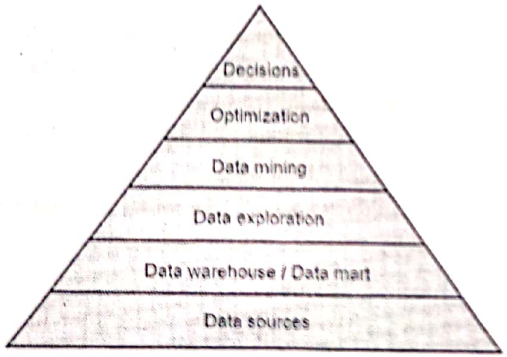


Fig. 1.8.1 Main components of a business intelligence system

- **Data sources :** This component of BI involves various forms of stored data. It's about taking the raw data and using software applications to create meaningful data sources that each division can use to positively impact business.

- BI analysts using this strategy may create data tools that allow data to be put into a large cache of spreadsheets, pie charts, tables or graphs that can be used for a variety of business purposes.
- **Data warehousing :** Data warehousing lets business leaders sift through subsets of data and examine interrelated components that can help drive business. Looking at sales data over several years can help improve product development or tailor seasonal offerings.
- Data warehousing can also be used to look at the statistics of business processes including how they relate to one another. For instance, business owners can compare shipping times in different facilities to look at which processes and teams work most efficiently.
- Data warehousing also involves storing huge amounts of data in ways that are beneficial to different divisions within the company.
- **Data exploration :** It uses tools for performing a passive business intelligence analysis, which consist of query and reporting systems, as well as statistical methods.
- **Data Mining :** It is active business intelligence methodologies, whose purpose is the extraction of information and knowledge from data. These include mathematical models for pattern recognition, machine learning and data mining techniques.
- **Optimization :** It allows to determine the best solution out of a set of alternative actions, which is usually fairly extensive and sometimes even infinite.
- **Decisions :** It corresponds to the choice and the actual adoption of a specific decision, and in some way represents the natural conclusion of the decision-making process.
- **Online Analytical Processing (OLAP) :** This component of BI allows executives to sort and select aggregates of data for strategic monitoring. With the help of specific software products, a certification in business intelligence helps business owners can use data to make adjustments to overall business processes.
- **Corporate Performance Management (CPM) :** This set of tools allows business leaders to look at the statistics of certain products or services. For instance, a fast food chain may analyze the sale of certain items and make local, regional and national modifications on menu board offerings as a result. The data could also be used to predict in which markets a new product may have the best success.
- **Real-time BI :** Using software applications, a business can respond to real-time trends in email, messaging systems or even digital displays. Because it's all in real-

time, an entrepreneur can announce special offers that take advantage of what's going on in the immediate.

1.8.2 Architecture of BI

- Fig. 1.8.2 shows architecture of BI. Architecture consists of five layers. These layers are data source, ETL (Extract-Transform-Load), data warehouse, end user and metadata layers.

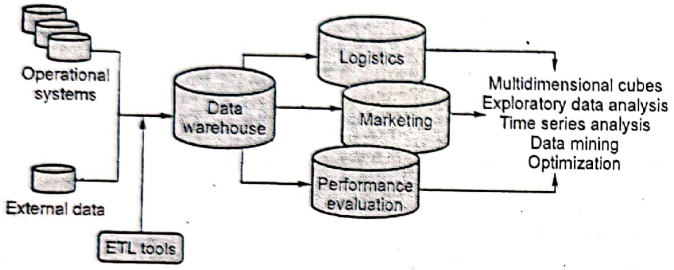


Fig. 1.8.2 Typical business intelligence architecture

- The five layers are data source, ETL (Extract-Transform-Load), data warehouse, end user, and metadata layers.

1. Data source layer :

- All data can be acquired from two types of sources : Internal and external.
- Internal data source refers to data that is captured and maintained by operational systems inside an organization such as Customer Relationship Management and Enterprise Resource Planning systems.
- Internal data sources include the data related to business operations.
- External data source refers to those that originate outside an organization. This type of data can be collected from external sources such as business partners, syndicate data suppliers, the Internet, governments and market research organizations .

2. ETL Layer

- This layer focuses on three main processes : extraction, transformation and loading
- Extraction is the process of identifying and collecting relevant data from different sources.

- The data collected from internal and external sources are not integrated, incomplete, and may be duplicated. Therefore, the extraction process is needed to select data that are significant in supporting organizational decision making.
- Transformation is the process of converting data using a set of business rules into consistent formats for reporting and analysis.
- Loading is the last phase of the ETL process. The data in staging area are loaded into target repository.
- Data warehouse layer : There are three components in the data warehouse layer: operational data store, data warehouse and data marts.
- Data flows from operational data store to data warehouse and subsequently to data mart.
- Operational Data Store (ODS) : ODS is used to integrate all data from the ETL layer and load them into data warehouses.

3. Data warehouse :

- Data warehouse is one of the most important components in BI architecture.
- It is defines data warehouse as "a subject-oriented, integrated, time-variant, and non-volatile collection of data in support of management's decision making process".

4. Metadata layer

- Metadata refers to data about data. It describes where data are being used and stored, the source of data, what changes have been made to the data, and how one piece of data relates to other information.
- Metadata repository is used to store technical and business information about data as well as business rules and data definitions.
- Good management and use of metadata can reduce development time, simplify on-going maintenance, and provide users with information about data source.

5. End user layer

- The end user layer consists of tools that display information in different formats to different users.
- These tools can be grouped hierarchically in a pyramid shape.

1.8.3 Cycle of a Business Intelligence Analysis

- Each business intelligence analysis follows its own path according to the application domain, the personal attitude of the decision makers and the available analytical methodologies.
- Fig. 1.8.3 shows cycle of a business intelligence analysis.

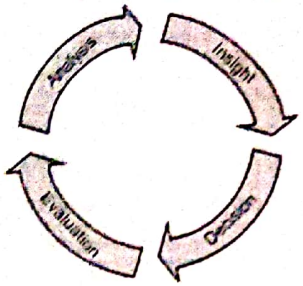


Fig. 1.8.3 Cycle of a business intelligence analysis

1. Analysis : When we gather data about our business we select only what we consider that is important. For example, we collect information to determine which factors affect our costs, our revenue or data that is important to our customers, to our partners or to our employees.

- We do this because we have a mental model of what our business is and how to influence in the results of our business. This mental model simplifies reality so we can make decisions easier and faster, not considering all the complexities that the real world has. Our mental model however, can also block us from seeing what might be obvious to others.

2. Insight : There are different types of insights. There are operational insights, like the cause of a decrease in sales of a certain product line. There are strategic insights, for example, the best strategy to penetrate into a new market category is through educating your target audience about your new product.

- Insight is the product of broad, "out of the box" thinking and analysis that only we can recognize as useful. If an individual has an important insight, it generally becomes useful when shared with others. Insights that bring change to mental models are usually resisted and sometimes unwelcome.

- Business intelligence leads us to the insights, but it also provides us with the data, patterns and logic to support our insights. It also helps us present the justification of our insights.

3. Decision : Business intelligence allows you to make better and faster decisions. Actions always follow these decisions. Decisions backed by good analysis and insights give courage to the action maker. To implement these decisions you will always need some strong organizational support. When actions are backed by strong analysis and business intelligence, the purpose of these actions is usually clearer and easily justified, hence making it easier to gain the support that you need.
4. Evaluation : Last phase of the business intelligence cycle involves performance measurement and evaluation. Extensive metrics should then be devised that are not exclusively limited to the financial aspects but also take into account the major performance indicators defined for the different company departments.

1.8.4 Enabling Factors in Business Intelligence Projects

- Enabling factors in business intelligence projects are technologies, analytics and human resources
- Technologies : Hardware and software technologies are significant enabling factors that have facilitated the development of business intelligence systems within enterprises and complex organizations.
- Analytics : Mathematical models and analytical methodologies play a key role in information enhancement and knowledge extraction from the data available inside most organizations.
- Human resources : The human assets of an organization are built up by the competencies of those who operate within its boundaries, whether as individuals or collectively.

1.8.5 Development of a Business Intelligence System

- Fig. 1.8.4 shows phases of business intelligence system development.
- Analysis : In this phase, requirement of the organization relative to the development of a business intelligence system should be carefully identified. Interview of knowledge workers is conducted who performing different roles in organizations.

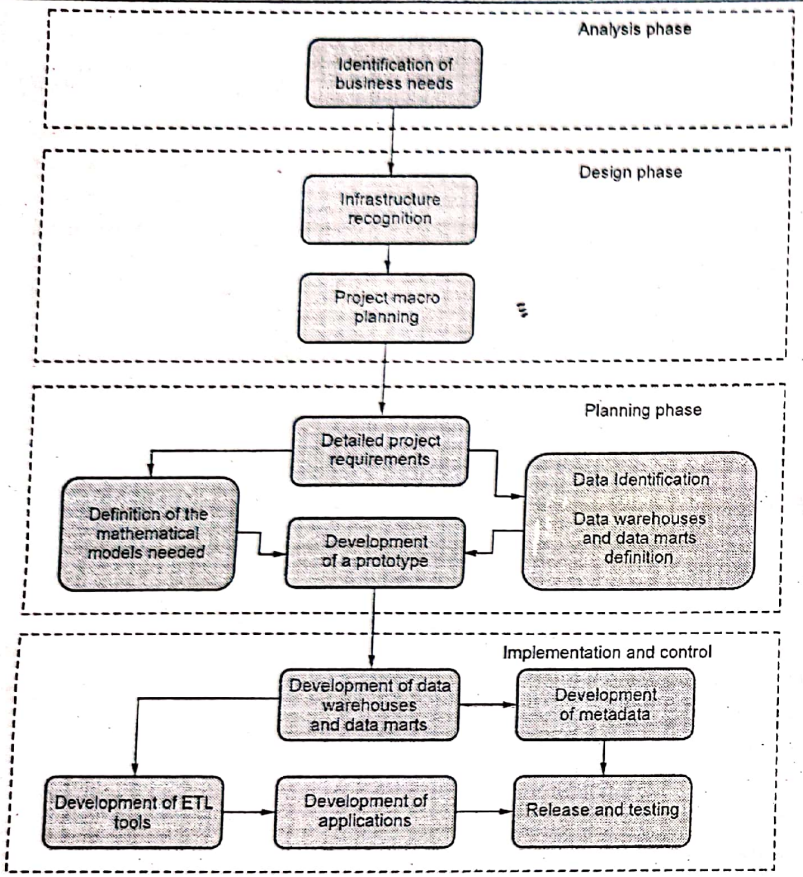


Fig. 1.8.4 Phases of business intelligence system development

- Design : It is necessary to make an assessment of the existing information infrastructures. The main decision-making processes that are to be supported by the business intelligence system should be examined, in order to adequately determine the information requirements. Later on, using classical project management methodologies, the project plan will be laid down, identifying development phases, priorities, expected execution times and costs, together with the required roles and resources.

- Planning : Existing data as well as other data that might be retrieved externally are assessed. It is appropriate to create a system prototype, at low cost and with limited capabilities.
- Implementation and control : It consists of five main sub-phases. These are as follows :
  a. Development of data warehouses and data marts
  b. Development of metadata
  c. Development of ETL tools
  d. Development of applications
  e. Release and testing
- The data warehouse and each specific data mart are developed. These represent the information infrastructures that will feed the business intelligence system.

1.8.6 Benefit of Business Intelligence

- Business intelligence is faster more accurate process of reporting critical information.
- BI facilitates better and efficient decision-making process.
- It provides timely information for better customer relationship management.
- Profitability of the company is improved.
- It provides a facility of assessing organization's readiness in meeting new business challenges.
- Business intelligence supports usage of best practices and identifies every hidden cost.

1.9 Previewing the Future of BI

- The future of business intelligence, therefore, lies in increasing the value and usability of this information. Business leaders are now able to make smart decisions, guided by expert, accurate data, to a level that is unprecedented.
- But business intelligence does not stop at data collection; it includes the transformation of this data, the presentation of data analysis to people within a business and the interpretation of this data when making plans or evaluating business processes.
- Today, BI provides organizations with new and novel ways to improve productivity, increase profits, and understand customers.

- *Data governance* : In the realm of BI, data governance will become a top priority for organizations big and small. This will also be driven by the number and complexity of data sources and data types needed to support analytics initiatives that are increasing exponentially. Data governance will enable organizations to clearly understand the information needs of the enterprise.
- *Self-service BI* : With centralized data from across the organization, users want to use the tool of their choice to drive value. Since business users know what they want, having a self-service data model will enable them to become self-sufficient. Self-service BI will break the dependency on IT/data teams to get access to the right data, and users will be able to meet their analytical requests with ease and make critical decisions at a much faster pace.
- *Prescriptive analytics* will examine data or content to determine what steps need to be taken to achieve an intended goal. By gauging the impact of future decisions, it will enable organizations to adjust the decisions before they are made – thereby improving decision-making accuracy.
- BI through Natural Language Processing (NLP) Using NLP, businesses will analyze customer sentiments, abstract information from a piece of text, or determine how positive (or negative) social media buzz around them is. The technology will also evolve into AI-powered chat bots, providing quick and accurate answers to users' BI inquiries.
- Business Intelligence-as-a-Service : BI-as-a-Service will enable organizations to get a BI solution up and running in no time while freeing their IT staff from carrying out complex analysis tasks. The model will allow organizations to get ready access to expert BI consultants and data architects who understand data and help manage.

1.10 Crafting a Better Experience for All Business Users

- Many executives think of business intelligence merely as a software solution that needs to be bought and installed, a reporting tool for serving up data on a convenient "dashboard." As a result of this misperception and despite the significant procurement, installation, and maintenance costs, BI systems often generate inaccurate data or distract employees by delving too deeply into corporate minutiae.
- A fairly small number of executives and companies, in contrast, have discovered that true business intelligence is the key to running a performance-oriented organization. They use their systems to home in on a selected group of key

performance indicators, often custom-crafted, that help them define corporate strategy and drive profitability.

- They have found that the data they receive gives them the ability to make sense of markets; to identify strengths and weaknesses; to measure the progress of the company against its goals; and to employ the skills, processes, technologies, applications and practices that support good decision making.
- One leading global logistics provider, for example, which had grown to more than 470,000 employees in 220 countries, recognized its need to reduce complexity, improve transparency and transition from intuition-based to fact-based decision making.
- The company designed a new common reporting system that consolidated four business units and more than 3,000 reporting entities worldwide. And the key performance indicators that emerged as a result enabled the management team to improve financial reporting capabilities, increase financial control and transparency throughout the company and harmonize the financial systems and saved huge amount.
- The reason that most companies are not getting the most out of their business intelligence has nothing to do with the software itself. Most off-the-shelf BI products are easy to implement and incredibly powerful; they are rich with features and capable of aggregating, integrating and analyzing data from nearly any part of the organization.
- The reason BI seems to be failing companies is that many of them have stumbled in their early attempts to leverage this performance-driven approach to running a business. Very few companies have the discipline to focus their operations in every business unit and product line on the things they do best. Those that do are the companies that identify their strongest internal capabilities; set thoughtful, strategic goals for them; and then constantly almost obsessively measure their performance against those goals.
- When deployed properly, business intelligence should help define strategy, drive profitability and develop a performance-oriented culture throughout an organization. It is much more than a reporting tool.
- Using BI is a way of doing business. Conceiving of metrics that will measure progress toward specific goals is critical. Once the right metrics have been

identified, executives should focus on gaining the support of key stakeholders and the cooperation of their employees and partners to ensure smooth implementation.

1.11 End User Assumptions

- The business intelligence end-user can be defined as a decision-maker, who does not necessarily possess IT skills and who uses business data and information from the BI solution to guide his actions.
- The true test of the usability of a BI solution is with the nontechnical end-user.
- The success that a BI solution will have in propelling the organization forward depends in large part on how it is received by end-users. Adoption makes or breaks a BI project. And adoption is, in turn, dependent on three factors : ease of use, usefulness and cost.
- Business intelligence goal is to help end-users solve problems, eliminate inefficiency and achieve the company's strategic goals. A well-implemented BI solution that is squarely and intelligently aligned with the company's strategy has indeed the potential to make a tremendous impact - if adopted.
- The first condition to adoption is ease of use. New technologies tend to make some new users anxious, especially if they are perceived as coming with a steep learning curve. If a newly-implemented BI solution is complex to learn and use, user can rest assured that many end-users will be reluctant to adopt it, and will instead fall back on what's familiar.
- Even if initially adopted, a BI solution will quickly lose its following within the organization if it does not provide real solutions for the end-users

1.11.1 Setting up Data for BI

- Preparing data for Business Intelligence can be a very tedious and time consuming process. User want the data to turn into the best reports for analysis. But, the raw data needs lots of processing and handling before user can even approach the results. In addition, it's essential to make sure data is collected and shared across the whole organization.
- Collecting data is an integral part of a business's success; it can enable user to ensure the data's accuracy, completeness and relevance to your organization and the issue at hand. The information gathered allows organizations to analyze past strategies and stay informed on what needs to change.

- In general, business analytics and reporting are enablers for business owners and leaders at all levels to change and direct their product. However, getting these results is not necessarily a straightforward endeavor. The data may contain lots of anomalies and duplication, it require redundancy removal, normalization across the different data sources, and varying granularity.
- It's also a challenge to get the data ready for everyone, not just the business owner and developers, but also for the key decision-makers. Infrastructure and tool challenges might actually slow down access to your data.
- To reach our desired results, there are several steps to take to go from raw data to useful analytics :
  1. Collect and load data
  2. Transform data to be BI ready
  3. Test system with manual queries
  4. Build the reports

1.12 Data, Information and Knowledge

- Fig. 1.12.1 shows relation between data, information and knowledge.

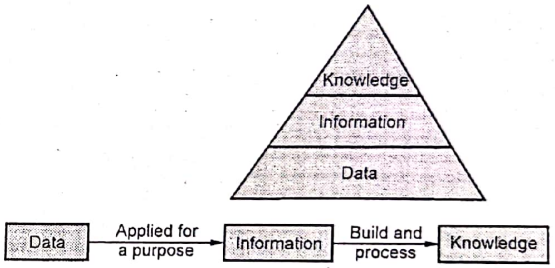


Fig. 1.12.1 Relation between data, information and knowledge

1. Data :

- Data is collection of facts and figures which relay something specific, but which are not organized in any way. It can be numbers, words, measurements, observations or even just descriptions of things. We can say, data is raw material in the production of information.
- **Examples of data :** Printed paper, bank account passbook, student attendance, salary sheet etc.

- Data creation is limited because of lack of new technology. After start of using computer, data can be converted into more convenient forms. It starts of using email, e-book, digital audio and video, images etc.
- Data is raw. It simply exists and has no significance beyond its existence. It can exist in any form, usable or not. It does not have meaning of itself. In computer parlance, a spreadsheet generally starts out by holding data.
- Facts and figures which relay something specific, but which are not organized in any way and which provide no further information regarding patterns, context, etc.
- Data represents a fact or statement of event without relation to other things. Ex : It is raining. Computer stores data in the form of 0 and 1. This is called digital data. These data is stored in the computer in the form of 0 and 1. Data in this form is called digital data.
- Data is raw. It has not been shaped, processed or interpreted. Once data has been processed and turned into information. Computers need data. Humans need information. Data is a building block. Information gives meaning and context

2. Information :

- Information is the outcome of extraction and processing activities carried out on data and it appears meaningful for those who receive it in a specific domain.
- Information is a subjective, meaningful interpretation of the data. Information has various purposes in creating knowledge, decision making and guiding further actions.

3. Knowledge :

- Knowledge is the collection of all that is known; the awareness or familiarity gained by experience, a person's range of information, a theoretical or practical understanding of a subject, language etc. In an information system, knowledge is the application of information by the use of rules.
- Information is transformed into knowledge when it is used to make decisions and develop the corresponding actions.
- The activity of providing support to knowledge workers through the integration of decision-making processes and enabling information technologies is usually referred to as knowledge management.

1.13 The Role of Mathematical Models

- A business intelligence system provides decision makers with information and knowledge extracted from data, through the application of mathematical models and algorithms.
- In some instances, this activity may reduce to calculations of totals and percentages, graphically represented by simple histograms, whereas more elaborate analyses require the development of advanced optimization and learning models.
- The adoption of a business intelligence system tends to promote a scientific and rational approach to the management of enterprises and complex organizations.
- Business intelligence analysis can be summarized schematically in the following main characteristics.
  1. The objectives of the analysis are identified and the performance indicators that will be used to evaluate alternative options are defined.
  2. Mathematical models are then developed by exploiting the relationships among system control variables, parameters and evaluation metrics.
  3. What-if analyses are carried out to evaluate the effects on the performance determined by variations in the control variables and changes in the parameters.

1.14 Ethics and Business Intelligence

- A review of the strategic management, policy, information management and the marketing literature reveals that many large and medium sized companies now collect and use business intelligence. The number of firms engaging in these activities is increasing rapidly.
- The adoption of business intelligence methodologies, data mining methods and decision support systems raises some ethical problems. Business ethics enhances the law by outlining acceptable behaviors beyond government control. Corporations establish business ethics to promote integrity among their employees and gain trust from key stakeholders, such as investors and consumers.
- Will business ethics play a part in any data analyses? Technology gives us the ability to analyze data in just about any way executives want. Should we analyze data in some ways, however? The "should we" aspect is the one that's often ignored and the real issue that privacy groups should focus on.

- The Internet plays a vital role in data collection, information creation and business intelligence (BI). The nature of information collected on the Internet and the degree to which such information is collected, both have ethical ramifications. What data can be collected is very different from what data should be collected.
- Usage of data by public and private organizations that is improper and does not respect the individuals' right to privacy should not be tolerated.

❑❑❑

Notes

Unit II

2 The Architecture of DW and BI

Syllabus

BI and DW architectures and its types - Relation between BI and DW - OLAP (Online analytical processing) definitions - Different OLAP Architectures-Data Models-Tools in Business Intelligence - Role of DSS, EIS, MIS and digital Dash boards - Need for Business Intelligence

Difference between OLAP and OLTP - Dimensional analysis - What are cubes ? Drill-down and roll-up - slice and dice or rotation - OLAP models - ROLAP versus MOLAP - defining schemas : Stars, snowflakes and fact constellations.

Contents

2.1 Data Warehousing Concept

- Data warehousing is the process of constructing and using a data warehouse. A data warehouse is constructed by integrating data from multiple heterogeneous sources that support analytical reporting, structured and/or ad hoc queries and decision making. Data warehousing involves data cleaning, data integration and data consolidations.
- A data warehouse is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision-making process. A data warehouse stores historical data for purposes of decision support.
- A database an application-oriented collection of data that is organized, structured, coherent, with minimum and controlled redundancy, which may be accessed by several users in due time.
- Data warehousing provides architectures and tools for business executives to systematically organize, understand and use their data to make strategic decisions.
- A data warehouse is a subject-oriented collection of data that is integrated, time-variant, non-volatile, which may be used to support the decision-making process.
- Data warehouses are databases that store and maintain analytical data separately from transaction-oriented databases for the purpose of decision support. Data warehouses separate analysis workload from transaction workload and enable an organization to consolidate data from several source.
- Data organization in data warehouses is based on areas of interest, on the major subjects of the organization : Customers, products, activities etc. databases organize data based on enterprise applications resulted from its functions.
- The main objective of a data warehouse is to support the decision-making system, focusing on the subjects of the organization. The objective of a database is to support the operational system and information is organized on applications and processes.
- A data warehouse usually stores many months or years of data to support historical analysis. The data in a data warehouse is typically loaded through an Extraction, Transformation and Loading (ETL) process from multiple data sources.
- Databases and data warehouses are related but not the same.
- A **database** is a way to record and access information from a single source. A database is often handling real-time data to support day-to-day business processes like transaction processing.

- A **data warehouse** is a way to store historical information from multiple sources to allow you to analyse and report on related data (e.g., your sales transaction data, mobile app data and CRM data). Unlike a database, the information isn't updated in real-time and is better for data analysis of broader trends.
- Modern data warehouses are moving toward an Extract, Load, Transformation (ELT) architecture in which all or most data transformation is performed on the database that hosts the data warehouse.
- Goals of data warehousing :
  1. To help reporting as well as analysis.
  2. Maintain the organization's historical information.
  3. Be the foundation for decision making.

"How are organizations using the information from data warehouses ?"

- Most of the organizations makes use of this information for taking business decision like :
  a) Increasing customer focus : It is possible by performing analysis of customer buying.
  b) Repositioning products and managing product portfolios by comparing the performance of last year sales.
  c) Analysing operations and looking for sources of profit.
  d) Managing customer relationships, making environmental corrections and managing the cost of corporate assets.

2.1.1 Characteristics of Data Warehouse

1. Subject oriented : Data are organized based on how the users refer to them. A data warehouse can be used to analyse a particular subject area. For example, "sales" can be a particular subject.
2. Integrated : All inconsistencies regarding naming convention and value representations are removed. For example, source A and source B may have different ways of identifying a product, but in a data warehouse, there will be only a single way of identifying a product.
3. Non-volatile : Data are stored in read-only format and do not change over time. Typical activities such as deletes, inserts and changes that are performed in an operational application environment are completely non-existent in a DW environment.

4. Time variant : Data are not current but normally time series. Historical information is kept in a data warehouse. For example, one can retrieve files from 3 months, 6 months, 12 months, or even previous data from a data warehouse.

Key characteristics of a Data Warehouse

1. Data is structured for simplicity of access and high-speed query performance.
2. End users are time-sensitive and desire speed-of-thought response times.
3. Large amounts of historical data are used.
4. Queries often retrieve large amounts of data, perhaps many thousands of rows.
5. Both predefined and ad hoc queries are common.
6. The data load involves multiple sources and transformations.

2.1.2 Multitier Architecture of Data Warehouse

- Data warehouse architecture is a data storage framework's design of an organization. A data warehouse architecture takes information from raw sets of data and stores it in a structured and easily digestible format.
- Data warehouse system is constructed in three ways. These approaches are classified the number of tiers in the architecture.
  a) Single-tier architecture.
  b) Two-tier architecture.
  c) Three-tier architecture (Multi-tier architecture).
- *Single tier* warehouse architecture focuses on creating a compact data set and minimizing the amount of data stored. While it is useful for removing redundancies. It is not effective for organizations with large data needs and multiple streams.
- *Two-tier* warehouse structures separate the resources physically available from the warehouse itself. This is most commonly used in small organizations where a server is used as a data mart. While it is more effective at storing and sorting data. Two-tier is not scalable and it supports a minimal number of end-users.

Three tier (Multi-tier) architecture :

- Three tier architecture creates a more structured flow for data from raw sets to actionable insights. It is the most widely used architecture for data warehouse systems.

- Fig. 2.1.1 shows three tier architecture. Three tier architecture sometimes called multi-tier architecture.

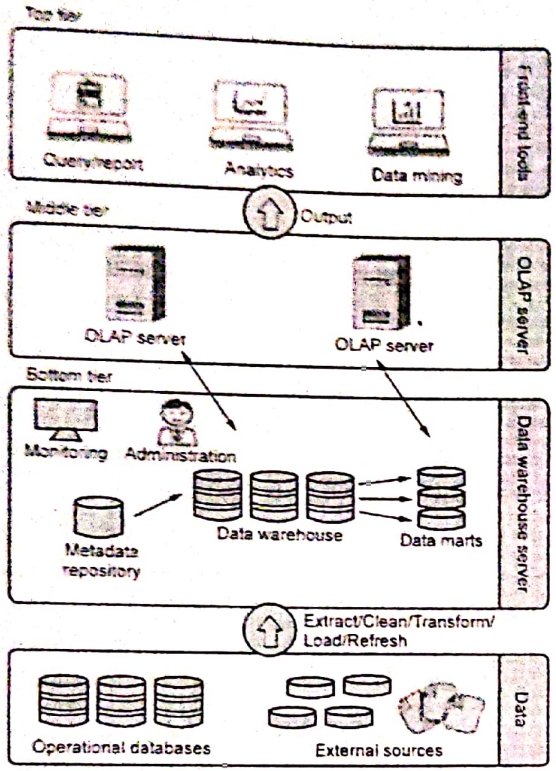


Fig. 2.1.1 Three tier architecture

- The bottom tier is the database of the warehouse, where the cleansed and transformed data is loaded. The bottom tier is a warehouse database server.
- The middle tier is the application layer giving an abstracted view of the database. It arranges the data to make it more suitable for analysis. This is done with an OLAP server, implemented using the ROLAP or MOLAP model.
- OLAPS can interact with both relational databases and multidimensional databases, which lets them collect data better based on broader parameters.

- The top tier is the front-end of an organization's overall business intelligence suite. The top-tier is where the user accesses and interacts with data via queries, data visualizations and data analytics tools.
- The top tier represents the front-end client layer. The client level which includes the tools and Application Programming Interface (API) used for high-level data analysis, inquiring and reporting. User can use reporting tools, query, analysis or data mining tools.

2.1.3 Needs of Data Warehouse

1) Business user : Business users require a data warehouse to view summarized data from the past. Since these people are non-technical, the data may be presented to them in an elementary form.
2) Store historical data : Data warehouse is required to store the time variable data from the past. This input is made to be used for various purposes.
3) Make strategic decisions : Some strategies may be depending upon the data in the data warehouse. So, data warehouse contributes to making strategic decisions.
4) For data consistency and quality : Bringing the data from different sources at a commonplace, the user can effectively undertake to bring the uniformity and consistency in data.
5) High response time : Data warehouse has to be ready for somewhat unexpected loads and types of queries, which demands a significant degree of flexibility and quick response time.

2.1.4 Benefits of Data Warehouse

a) Understand business trends and make better forecasting decisions.
b) Data warehouses are designed to perform well enormous amounts of data.
c) The structure of data warehouses is more accessible for end-users to navigate, understand and query.
d) Queries that would be complex in many normalized databases could be easier to build and maintain in data warehouses.
e) Data warehousing is an efficient method to manage demand for lots of information from lots of users.
f) Data warehousing provide the capabilities to analyze a large amount of historical data.

2.1.5 Difference between ODS and Data Warehouse

Sr. No.	Operational data store	Data warehouse
1.	ODS uses current data.	Data warehouse uses historical data.
2.	Run the business on a current basis.	Support managerial decision making.
3.	Design goal is performance throughput and administrators.	Design goal is easy reporting and analytics.
4.	Frequent small updates.	Period batch updates.
5.	ODS does not support summary data.	Data warehouse support summary data.
6.	Supports simple queries on a few rows.	Supports complex queries on several rows.

2.1.6 Metadata

- Metadata is simply defined as data about data. The data that is used to represent other data is known as metadata. In data warehousing, metadata is one of the essential aspects.
- We can define metadata as follows :
  a) Metadata is the road-map to a data warehouse.
  b) Metadata in a data warehouse defines the warehouse objects.
  c) Metadata acts as a directory. This directory helps the decision support system to locate the contents of a data warehouse.
- In a data warehouse, we create metadata for the data names and definitions of a given data warehouse. Along with this metadata, additional metadata is also created for time-stamping any extracted data, the source of extracted data.

Why is metadata necessary in a data warehouse ?

a) First, it acts as the glue that links all parts of the data warehouses.
b) Next, it provides information about the contents and structures to the developers.
c) Finally, it opens the doors to the end-users and makes the contents recognizable in their terms.

- Fig. 2.1.2 shows warehouse metadata.

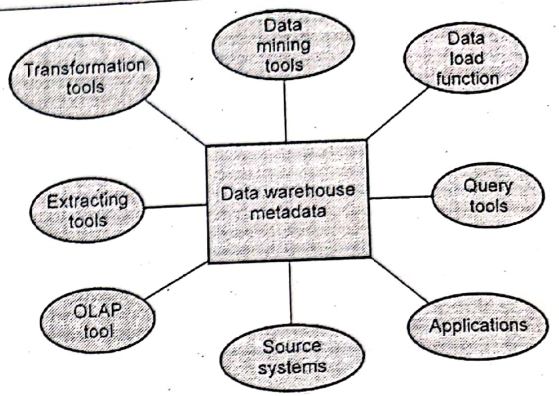


Fig. 2.1.2

Review Questions

1. *What is data warehouse ? Differentiate between ODS and data warehouse.*
2. *Explain with diagram a three-tier data warehouse architecture.*
3. *What is metadata in data warehouse ? What it contains ?*
4. *What is data warehouse ? Discuss various usage trends in data warehousing.*
5. *Explain in detail the three - tier data warehouse architecture.*

2.2 Data Warehouse Types

- In a traditional architecture, there are three common data warehouse models : The enterprise warehouse, the data mart and the virtual warehouse.

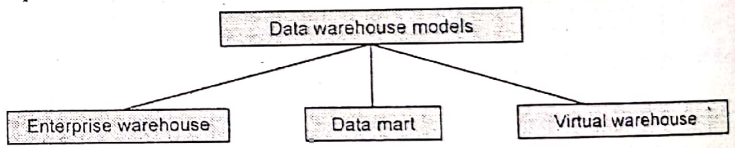


Fig. 2.2.1

1. **Virtual warehouse :**
   - A virtual warehouse is a set of views over operational databases. For efficient query processing, only some of the possible summary views may be materialized.
   - A virtual warehouse is easy to build but requires excess capacity on operational database servers.

- A virtual data warehouse is a set of separate databases, which can be queried together, so a user can effectively access all the data as if it was stored in one data warehouse.

2. **Data mart :**
   - A data mart model is used for business-line specific reporting and analysis. In this data warehouse model, data is aggregated from a range of source systems relevant to a specific business area, such as sales or finance.
   - A data mart is a lower-cost, scaled-down version of a data warehouse, usually designed to support a small group of users.
   - Data mart offer a targeted and less costly method of gaining the advantages associated with data warehousing and can be scaled up to a full DW environment over time.
   - Depending on the source of data, data marts can be categorized as independent or dependent.
   - Independent data marts are sourced from data captured from one or more operational systems or external information providers or from data generated locally within a particular department or geographic area.
   - Dependent data marts are sourced directly from enterprise data warehouses.
3. **Enterprise warehouse :**
   - An enterprise warehouse collects all of the information about subjects spanning the entire organization.
   - An enterprise data warehouse model prescribes that the data warehouse contains aggregated data that spans the entire organization. This model sees the data warehouse as the heart of the enterprise's information system, with integrated data from all business units.
   - It provides corporate-wide data integration, usually from one or more operational systems or external information providers and is cross-functional in scope

2.2.1 Extraction, Transformation and Loading

- **Extraction, Transformation and Loading (ETL)** is a process of data integration that encompasses three steps : extraction, transformation and loading. Fig. 2.2.2 shows ETL process.

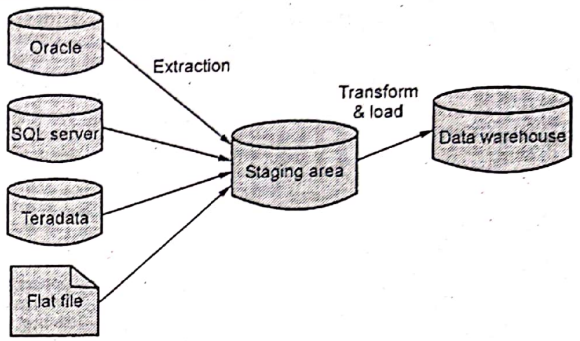


Fig. 2.2.2 ETL process

- ETL is a data integration methodology that extracts raw data from sources, transforms the data on a secondary processing server and then loads the data into a target database.
- Extraction is the process of identifying and collecting relevant data from different sources. Data extraction which typically gathers data from multiple, heterogeneous and external sources. The staging area acts as a buffer between the data warehouse and the source data. The staging area is used for data cleansing and organization.
- Transformation : The data cleaning and organization stage is the transformation stage. The data collected from internal and external sources are not integrated, incomplete and may be duplicated. Therefore, the extraction process is needed to select data that are significant in supporting organizational decision making.
- Transformation is the process of converting data using a set of business rules into consistent formats for reporting and analysis.
- Loading is the last phase of the ETL process. The data in staging area are loaded into target repository. All the gathered information is loaded into the target data warehouse tables.

2.3 Relation between BI and DW

Business Intelligence	Data Warehouse
Business Intelligence is a set of strategies and technologies to analyze and visualize data to make business decisions.	Data Warehouse is a central location that is used to store consolidated data from multiple data sources

The department store in the beginning uses data to predict that the user's shopping behavior belongs to business intelligence	The customer's consumption habits are stored in the data warehouse.
Business intelligence is a set of technologies and strategies.	Data warehouse is a storage
BI presents data as reports, charts and graphs.	Data warehouse presents data in tables
Top executives and senior managers use business intelligence.	Data Engineers, data and business analysts use data warehouses
Example : Datapine	Example : Amazon Redshift

2.4 Online Analytical Processing (OLAP)

- Online Analytical Processing (OLAP) is a category of software that allows users to analyse information from multiple database systems at the same time. It is a technology that enables analysts to extract and view business data from different points of view.
- As online analytical processing operations is a multidimensional data model, these operations are performed over the data cubes. The concept of data cube is used because to represent the data in three-dimensional space and so that analysts can turn around the data cube along its dimensions for all its possible space combinations to determine the every aspect of available data.
- Each OLAP cube is presented through measures and dimensions. Measures refers to the numeric value categorized by dimensions.
- Basic operations are roll-up (consolidation), drill-down, slicing and dicing

1. Roll-up :

- It is also called drill-up operation. It performs aggregation on a data cube, either by climbing up a concept hierarchy for a dimension or by dimension reduction. Roll-up is like zooming-out on the data cube.
- The roll-up operation shown aggregates the data by ascending the location hierarchy from the level of city to the level of country.
- When roll-up is performed by dimension reduction, one or more dimensions are removed from the given cube.
- A roll-up involves summarizing the data along a dimension. The roll-up operation is performed by climbing up a concept hierarchy for the dimension location.

- When roll-up operation is performed then one or more dimensions from the data cube are removed.
- For example, consider a sales data cube containing only the two dimensions location and time. Roll-up may be performed by removing, say, the time dimension, resulting in an aggregation of the total sales by location, rather than by location and by time.
- Fig. 2.4.1 shows roll-up and drill-down operation.

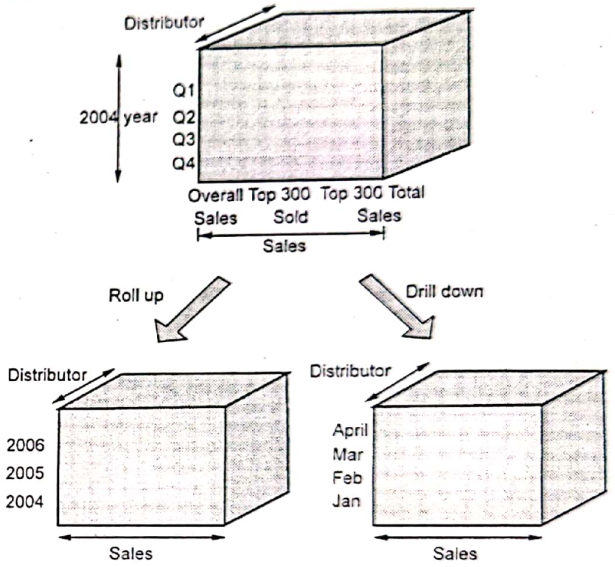


Fig. 2.4.1 Roll-up and drill-down operation

2. Drill-down

- Drill down is the reverse of roll-up.
- Drill down is a dimension expansion technique that can be applied on the data cube. Dimension expansion means, adding new dimension or expanding existing dimensions across any axis of the data cube using the notion of concept hierarchy.
- Navigates from less detailed data to more detailed data it can achieved.

3. Slice and dice :

- The slice operation performs a selection on one dimension of the given cube, resulting in a sub-cube.

- The dice operation defines a sub-cube by performing a selection on two or more dimensions.
- The slice operation produces a sliced OLAP cube by allowing the analyst to pick specific value for one of the dimensions.
- A slice in a multidimensional array is a column of data corresponding to a single value for one or more members of the dimension. It helps the user to visualize and gather the information specific to a dimension. When you think of slicing, think of it as a specialized filter for a particular value in a dimension.

4. Pivot (or rotate)

- Rotates the data axis to view the data from different perspectives.
- Groups data with different dimensions. Fig. 2.4.2 shows OLAP pivot operations.

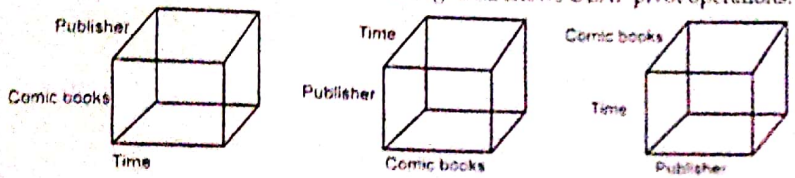


Fig. 2.4.2

2.4.1 Indexing OLAP Data : Bitmap Index and Join Index

- Indexes are database objects associated with database tables and created to speed up access to data within the tables.
- Indexing techniques have already been in existence for decades for the OLTP relational database system but they cannot handle large volume of data and complex and iterative queries that are common in OLAP applications.
- Indexes are data structures which hold field values from the indexed column(s) and pointers to the related record(s). This data structure is then sorted and binary searches are performed to quickly find the record.
- Data warehouse is a large repository of information accessed through an OLAP application. This application provides users with tools to iteratively query the DW in order to make better and faster decisions.
- The information stored in a DW is clean, static, integrated, and time varying and is obtained through many different sources.
- Requests for information from a DW are usually complex and iterative queries. Complex queries could take several hours or days to process because the queries

have to process through a large amount of data. So index structure is used in data warehouse.

To index OLAP data by bitmap indexing and join indexing

1. Bitmap indexing :

- Bitmap indexes are widely used in data warehousing environments. It allows quick searching in data cubes.
- Bitmap indexing enriches the implementation of individual query operators. With the help of bitmaps, query operators, such as selection, aggregate and Group By, and Join, are re-implemented.
- Index on a particular column. Each value in the column has a bit vector. The length of the bit vector is number of records in the base table.
- Following is Indexing OLAP data using bitmap indices.

Base Table		
Customer	Region	Type
C1	Asia	Retail
C2	UK	Dealer
C3	Asia	Dealer
C4	US	Retail
C5	UK	Dealer

Index on Region			
RecID	Asia	UK	US
1	1	0	0
2	0	1	0
3	1	0	0
4	0	0	1
5	0	1	0

Index on Type		
RecID	Retail	Dealer
1	1	0
2	0	1
3	0	1
4	1	0
5	0	1

- In a Bitmap Index, each distinct value for the specified column is associated with a bitmap where each bit represents a row in the table. A '1' means that row contains that value, a '0' means it doesn't.
- For given attribute, there is a distinct bit vector (Bv) for each value v in the domain of the attribute. If domain contains n value attributes, then n bits are required for each entry in the bitmap index.
- **Advantage of bitmap indexes :**
  a. They have a highly compressed structure, making them fast to read.
  b. Structure makes it possible for the system to combine multiple indexes together for fast access to the underlying table.
  c. Only suitable for indexing low - cardinality data.
  d. Reduces processing time.
- **Disadvantage of bitmap indexes :**
  a. The overhead on maintaining them is huge.

b. Modification to a bitmap index requires a great deal more work on behalf of the system than a modification to a b - tree index.

2. Join indexing :

- The join indexing method used in relational database query processing. Join indexing registers the joinable rows of two relations from a relational database.
- The join index records can identify joinable tuples without performing costly join operations. Join indexing is especially useful for maintaining the relationship between a foreign key and its matching primary keys, from the joinable relation.
- Join index is extremely useful for table joins that involve low - cardinality columns. A join index connect dimension data to tuples in a fact table.
- Fig. 2.4.3 shows join indexing.

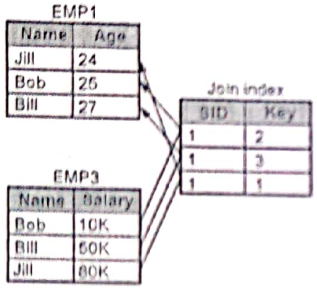


Fig. 2.4.3 Join indexing

- A bitmap join index can improve the performance by an order of magnitude. By storing the result of a join, the join can be avoided completely for SQL statements using a bitmap join index.
- Join indexes have the following restrictions :
  1. Only one table can be updated concurrently by different transactions when using the bitmap join index.
  2. No table can appear twice in the join.
  3. You cannot create a bitmap join index on a temporary table.
  4. The columns in the index must all be columns of the dimension tables.
  5. The dimension table join columns must be either primary key columns or have unique constraints.

2.4.2 Efficient Processing of OLAP Queries

- OLAP query processing is as follows :
  a) Find out which operations should be performed on the available cuboids. It is related to transforming selection, projection, roll - up and drill - down operations specified in the query into corresponding SQL and/or OLAP operations.

  Example : Slicing and dicing operation on a data cube may correspond to selection and/or projection operations on a materialized cuboid.

  b) Find out which materialized cuboid(s) the relevant operations should be applied. It identifies all of the materialized cuboids that may potentially be used to answer the query.
- **Why run OLAP queries over data warehouse ?** Warehouse collects and combines data from multiple sources. Warehouse may organize the data in certain formats to support OLAP queries.
- OLAP queries are complex and touch large amounts of data. They may lock the database for long periods of time and negatively affects all other OLTP transactions.

2.4.3 Role of Concept Hierarchies

- Concept hierarchy defines a sequence of mappings from a set of low-level concepts to higher-level. Many concept hierarchies are implicit within the database scheme.
- Fig. 2.4.4 show concept of hierarchy.

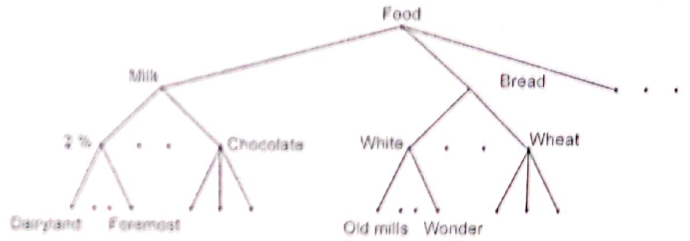


Fig. 2.4.4 Concept of hierarchy

- Categorical data are discrete data. Categorical attributes have a finite number of distinct values, with no ordering among the values.
- A concept hierarchy that is a total or partial order among attributes in a database schema is called a schema hierarchy.
- Example : Geographic location, job category and item type.

- Example : Suppose a user selects a set of location-oriented attributes, street, country, state and city, from the database, but does not specify the hierarchical ordering among the attributes.
- Fig. 2.4.5 shows automatic generation of a schema concept hierarchy based on the number of distinct attribute values.
- Various methods are used for the generation of concept hierarchies for categorical data

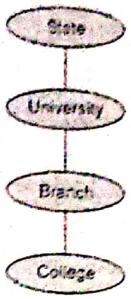


Fig. 2.4.5 Automatic generation of a schema concept hierarchy

a. Specification of a partial ordering of attributes explicitly at the schema level by users or experts

- Example : A relational database or a dimension location of a data warehouse may contain the following group of attributes : Street, city, province or state and country
- A user or expert can easily define a concept hierarchy by specifying ordering of the attributes at the schema level.
- A hierarchy can be defined by specifying the total ordering among these attributes at the schema level, such as : *street < city < province or state < country.*

b. Specification of a portion of a hierarchy by explicit data grouping

- We can easily specify explicit groupings for a small portion of intermediate-level data.
- For example, after specifying that area and country form a hierarchy at the schema level, a user could define some intermediate levels manually, such as : {India, Maharashtra, Pune} < SPPU.

c. Specification of a set of attributes, but not of their partial ordering

- A user may specify a set of attributes forming a concept hierarchy, but omit to explicitly state their partial ordering.
- The system can then try to automatically generate the attribute ordering so as to construct a meaningful concept.

Data discretization :

- Data discretization means diving the range of continuous attribute into intervals. Actual data values are replaced by interval labels.
- It reduces the number of values for a given continuous attribute. Some classification algorithms only accept categorical attributes. It helps to a concise, easy-to-use, knowledge-level representation of mining results.

- Data discretization techniques can be categorized based on class information and which direction it proceeds. Class information is divided into two types : Supervised and unsupervised discretization. Categorized based on which direction it proceeds are of two types : Top-down and bottom-up.
- Discretization techniques can be classified as supervised and unsupervised discretization. Supervised discretization uses class information and unsupervised discretization does not use class information.
- **Top-down :** If the process starts by first finding one or a few points to split the entire attribute range and then repeats this recursively on the resulting intervals. It is also called splitting.
- **Bottom-up :** It starts by considering all of the continuous values as potential split-points, removes some by merging neighbourhood values to form intervals and then recursively applies this process to the resulting intervals. It is also called merging.
- Discretization can be performed recursively on an attribute to provide a hierarchical or multiresolution partitioning of the attribute values, known as a concept hierarchy. Concept hierarchies are useful for mining at multiple levels of abstraction.
- Discretization and concept hierarchy generation for numerical Data uses following methods :

a. Binning b. Histogram analysis
c. Clustering analysis d. Entropy-based discretization
e. Segmentation by natural partitioning.

2.5 Different OLAP Architectures

- All OLAP architectures involve building a multidimensional data structure, where dimensions represent business entities such as sales regions and products or natural entities such as time and geography.
- OLAP is a database technology that has been optimized for querying and reporting, instead of processing transactions. OLAP data is also organized hierarchically and stored in cubes instead of tables.
- OLAP server aggregated data calculations and storage are performed by the server. It can process large amounts of data with multiple users.

- The multidimensional structure is organized in such a way that each data element is located and accessible based on the intersection of the dimension elements that define this element. Fig. 2.5.1 OLAP operation.

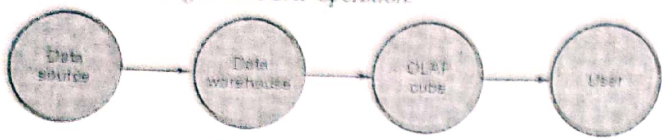


Fig. 2.5.1 OLAP operation

- OLAP server can operate the processed multidimensional information to provide users with consistent and fast response. The server can also populate its data structures in real time from different databases.
- OLAP databases contain two basic types of data : **Measures and dimensions.**
- Measures are numeric data, the quantities and averages that you use to make informed business decisions, and dimensions, which are the categories that you use to organize these measures. OLAP databases help organize data by many levels of detail, using the same categories that you are familiar with to analyze the data.
- An OLAP cube is a data structure that allows fast analysis of data. The arrangement of data into cubes overcomes a limitation of relational databases. It consists of numeric facts called measures which are categorized by dimensions.

Benefits of OLAP :

- One main benefit of OLAP is consistency of information and calculations.
- "What if" scenarios are some of the most popular uses of OLAP software and are made eminently more possible by multidimensional processing.
- It allows a manager to pull down data from an OLAP database in broad or specific terms.
- OLAP creates a single platform for all the information and business needs, planning, budgeting, forecasting, reporting and analysis.

2.5.1 Multidimensional OLAP

- Multidimensional OLAP (MOLAP) is the 'classic' form of OLAP and is sometimes referred to as just OLAP.
- MOLAP stores this data in an optimized multi - dimensional array storage, rather than in a relational database. Therefore it requires the pre - computation and storage of information in the cube - the operation known as **processing.**
- MOLAP tools generally utilize a pre - calculated data set referred to as a data cube.

- The data cube contains all the possible answers to a given range of questions.
- MOLAP tools have a very fast response time and the ability to quickly write back data into the data set.

Advantages of MOLAP :

1. **Excellent performance :** MOLAP cubes are built for fast data retrieval, and are optimal for slicing and dicing operations.
2. **Can perform complex calculations :** All calculations have been pre - generated when the cube is created.

Disadvantages of MOLAP :

- Limited in the amount of data it can handle.
- Requires additional investment : Cube technology are often proprietary and do not already exist in the organization.

2.5.2 Relational OLAP

- Relational OLAP (ROLAP) is a kind of online analytical processing (OLAP) that analyses data using multidimensional data models.
- ROLAP can handle large volumes of data.
- Although its use of a relational database means that it requires more processing time, it can be accessed by any SQL tool, and it does not have to be a tool specifically for OLAPs.
- Compared to MOLAP, ROLAP tools are much better at controlling non-aggregated facts such as textual descriptions.

Advantages of ROLAP :

1. It can handle large amounts of data : The data size limitation of ROLAP technology is the limitation on data size of the underlying relational database. In other words, ROLAP itself places no limitation on data amount.
2. Can leverage functionalities inherent in the relational database : Often, relational database already comes with a host of functionalities.

Disadvantages of ROLAP :

1. Performance can be slow : Because each ROLAP report is essentially a SQL query in the relational database, the query time can be long if the underlying data size is large.

2. Limited by SQL functionalities : Because ROLAP technology mainly relies on generating SQL statements to query the relational database, and SQL statements do not fit all needs.

2.5.3 Hybrid OLAP

- There is no clear agreement across the industry as to what constitutes Hybrid OLAP, except that a database will divide data between relational and specialized storage.
- For example, HOLAP database will use relational tables to hold the larger quantities of detailed data, and use specialized storage for at least some aspects of the smaller quantities of more - aggregate or less - detailed data.
- HOLAP addresses the shortcomings of MOLAP and ROLAP by combining the capabilities of both approaches.
- HOLAP tools can utilize both pre - calculated cubes and relational data sources

2.5.4 Comparison between ROLAP, MOLAP and HOLAP

Parameters	ROLAP	MOLAP	HOLAP
Implementation	An implementation based on relational DBMSs	An implementation based on multidimensional DBMSs	An implementation using both relational and multidimensional techniques
Advantages	• It can handle large amounts of data • Can leverage • Functionalities • inherent in the • relational database	• Excellent performance • Can perform complex calculations	• HOLAP tools can utilize both pre - calculated cubes and relational data sources
Disadvantages	• Performance can be • Slow • Limited by SQL functionalities	• Limited in the amount of data it can handle • Requires additional investment	• HOLAP supports disadvantages of MOLAP

2.5.5 Comparison between ROLAP and MOLAP

Characteristics	ROLAP	MOLAP
SCHEMA	• User star Schema. • Additional dimensions can be added dynamically.	• User Data cubes. • Addition dimensions require recreation of data cube.

Database Size	Medium to large	Small to medium
Architecture	Client/Server	Client/Server
Access	Support ad-hoc requests	Limited to pre - defined dimensions
Speed	• Good with small data sets. • Average for medium to large data set.	• Faster for small to medium data sets. • Average for large data sets.
Flexibility and Scalability	High	Low

2.5.6 Difference between OLAP and OLTP

Online Analytical Processing (OLAP)	Online Transaction Processing (OLTP)
It supports short transaction both query and updates.	It supports long transactions, usually complex queries.
An OLAP database has a multi-dimensional schema.	OLTP uses a traditional DBMS to accommodate a large volume of real-time transactions.
Tables in OLAP database are not normalized.	Tables in OLTP database are normalized.
OLAP systems are designed for use by data scientists, business analysts and knowledge workers.	OLTP systems are designed for use by frontline workers (e.g., cashiers, bank tellers, hotel desk clerks) or for customer self - service applications.
OLAP contains historical data.	OLTP contains current data.
It focuses on information out.	It is application oriented.
DB size : 100 MB to GB	DB size : 100 GB to TB
Used for decision support system.	Used for day to day operations.

Review Questions

1. *Explain OLAP operations with examples.*
2. *Explain indexing OLAP data : Bitmap index and join index with example.*
3. *Explain ROLAP versus MOLAP.*
4. *Summarize the various OLAP operations in the multidimensional data model.*
5. *Identify the different indexing method used for OLAP data with brief explanation.*
6. *Differentiate ROLAP, MOLAP and HOLAP servers.*

2.6 Data Models

- Data modeling is the process of creating a visual representation of either a whole information system or parts of it to communicate connections between data points and structures.
- The goal is to illustrate the types of data used and stored within the system, the relationships among these data types, the ways the data can be grouped and organized and its formats and attributes.
- Data models are built around business needs. Rules and requirements are defined upfront through feedback from business stakeholders so they can be incorporated into the design of a new system or adapted in the iteration of an existing one.
- Data can be modeled at various levels of abstraction. The process begins by collecting information about business requirements from stakeholders and end users. These business rules are then translated into data structures to formulate a concrete database design.
- A data model can be compared to a roadmap, an architect's blueprint or any formal diagram that facilitates a deeper understanding of what is being designed.
- Data modeling employs standardized schemas and formal techniques. This provides a common, consistent and predictable way of defining and managing data resources across an organization, or even beyond.
- Ideally, data models are living documents that evolve along with changing business needs. They play an important role in supporting business processes and planning IT architecture and strategy. Data models can be shared with vendors, partners, and/or industry peers.
- Types of data models are conceptual data models, logical data model and physical data models.
  a) Conceptual models are usually created as part of the process of gathering initial project requirements.
  b) Logical data models don't specify any technical system requirements.
  c) Physical data models can include database management system-specific properties, including performance tuning.

2.7 Tools in Business Intelligence

- Business Intelligence tools are proprietary or open source application software that are used to collect, process, analyze, sort, filter and report large quantities of data

from internal and external systems, for the purpose of transforming raw data into useful information for business purposes.

- Business Intelligence leverages BI reporting tools to transform data into actionable insights that improve business decisions. BI tools process and prepare raw data for analysis and facilitate the generation of reports, data visualizations and dashboards.
- BI data visualization tools aim to streamline the analysis process so that the average user can visualize, understand and draw conclusions from their data. The results empower businesses to inform and accelerate decision making, identify trends and revenue potential, increase efficiency, determine KPIs and reveal business opportunities.
- **Types of Business Intelligence Tools :** Common business intelligence tools include Dashboards, Visualizations, Reporting, Data mining, ETL and OLAP.
- Features of Business Intelligence tools :
  1) Spreadsheets : Organizes data in a tabular format that can easily be queried and formatted; available in web based format and downloaded software format.
  2) Dashboards : A real-time user interface that displays data visualizations that reflect the current status of data.
  3) Data mining tools : Data mining employs AI, Machine Learning, statistics, and database systems to reveal patterns in data.
  4) Ad hoc data analytics : An analysis process designed to answer specific questions on the spot
  5) Online analytical processing (OLAP) : OLAP business intelligence tools provide a computing method that enables multi-dimensional analytical queries
  6) Mobile BI : Software that optimizes desktop business intelligence for mobile devices
  7) Real-time BI : An advanced enterprise analytics approach that delivers real-time information to users by feeding business transactions into a real-time data warehouse
  8) Operational BI : A data analysis approach that utilizes real-time business analytics to automatically integrate real-time data into an operational system for immediate use
  9) Open source BI (OSBI) : Business intelligence software solutions that do not require purchasing a software license

10) Collaborative BI : The merging of business intelligence software with collaboration tools in order to streamline the sharing process

11) Data visualization software : Facilitates the detection of patterns and correlations by providing visual context

12) Reporting and query software : Functions to report, query, sort, filter and display data

13) Data warehousing tools : Integrated data retrieval from different sources creates a consolidated repository for data storage, which can be retrieved in the future for analysis

2.8 Data Cube

- A data cube allows data to be modelled and viewed in multiple dimensions. It is defined by dimensions and facts. A data cube in a data warehouse is a multidimensional structure used to store data.
- Data cube represents the data in terms of dimensions and facts. A data cube is used to represents the aggregated data. A data cube is basically categorized into two main kinds that are multidimensional data cube and relational data cube.
- In data warehousing literature, an n-D base cube is called a base cuboid. The top most 0-D cuboid, which holds the highest-level of summarization, is called the apex cuboid. The lattice of cuboids forms a data cube. Each node in the lattice represents one possible grouping/aggregation.

Example 2.8.1 *Suppose that a data warehouse consists of the three dimensions time, doctor, and patient, and the two measures count and charge, where charge is the fee that a doctor charges a patient for a visit. Draw the lattice of cuboids (from apex to base cuboid) for the given data warehouse.*

Solution : Lattice of cuboids for the given data warehouse

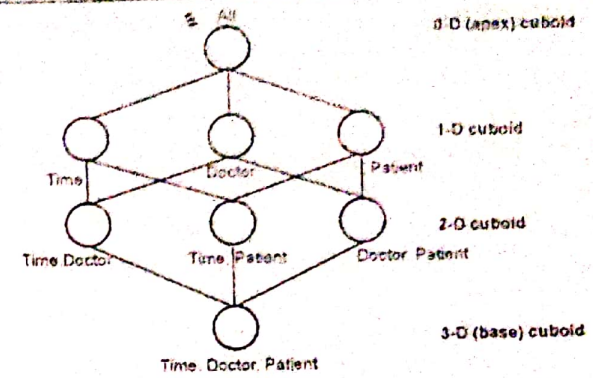


Fig. 2.8.1

- At the core of multidimensional data analysis is the efficient computation of aggregations across many sets of dimensions. In SQL terms, these aggregations are referred to as group-by's. Each group-by can be represented by a cuboid, where the set of group-by's forms a lattice of cuboids defining a data cube.
- The compute cube operator computes aggregates over all subsets of the dimensions specified in the operation.

Curse of Dimensionality :

- Dimensionality in statistics refers to how many attributes a dataset has.
- The curse of dimensionality usually refers to what happens when you add more and more variables to a multivariate model. The more dimensions you add to a data set, the more difficult it becomes to predict certain quantities. However, when it comes to adding variables, the opposite is true. Each added variable results in an exponential decrease in predictive power. This problem is referred to as the **curse of dimensionality.**
- "How many cuboids are there in an n-dimensional data cube?" If there were no hierarchies associated with each dimension, then the total number of cuboids for an n-dimensional data cube, as we have seen, is 2^n.

Data Cube Materialization :

- Data cube materialization for base cuboid are as follows :
  1. **No materialization :** Do not pre-compute any of the "non-base" cuboids. This leads to computing expensive multidimensional aggregates on the fly, which can be extremely slow.

2. **Full materialization :** Pre-compute all of the cuboids. The resulting lattice of computed cuboids is referred to as the full cube. This choice typically requires huge amounts of memory space in order to store all of the pre-computed cuboids.
3. **Partial materialization :** Selectively compute a proper subset of the whole set of possible cuboids.

2.8.1 Advantages of Data Cube

- Data cube ease in aggregating and summarizing the data.
- Data cube provide better visualization of data.
- Data cube stores huge amount of data in a very simplified way.
- Data cube increases the overall efficiency of the data warehouse.
- The aggregated data in data cube helps in analysing the data fast and thereby reducing the access time.

2.9 Defining Schemas

- A multi-dimensional data model is logical view of an enterprise that represents the important entities of a business and the relationship between them. It is not restricted to a physical database and tables. It's not represented by E-R diagrams.
- The multidimensional data model holds data in the shape of a data cube. Two or three-dimensional cubes are often served by data warehousing.
- Multidimensional modelling is a technique for structuring data around the business concepts. ER models describe "entities" and "relationships". Multidimensional model describes "measures" and "dimensions".
- A **data cube** allows data to be modelled and viewed in multiple dimensions. It is defined by dimensions and facts.
- Dimensions are the perspectives or entities with respect to which an organization wants to keep records. Each dimension may have a table associated with it, called a dimension table.
- The cube is used to represent data along some measure of interest. Although called a "cube", it can be 2-dimensional, 3-dimensional or higher-dimensional. Each dimension represents some attribute in the database and the cells in the data cube represent the measure of interest.
- For example, they could contain a count for the number of times that attribute combination occurs in the database or the minimum, maximum, sum or average value of some attribute. Queries are performed on the cube to retrieve decision support information.

- A multidimensional database allows to rapidly and reliably providing data-related responses to complicated market questions. The multidimensional data model can be defined as a way to arrange the data in the database, to help structure and organize the contents of the database.
- The multidimensional data model can include two or three dimensions of objects from the database structure, versus a system of one dimension, such as a list.
- Data cube allows data to be modeled and viewed in multiple dimensions. It is defined by dimensions and facts.
- OLAP is based on a multidimensional data model which views data in the form of a data cube. A data cube, such as sales, allows data to be modeled and viewed in multiple dimensions.
- For example, All product may create a sales data warehouse in order to keep records of the store's sales with respect to the dimensions time, item, branch and location.
- These dimensions allow the store to keep track of things like monthly sales of items and the branches and locations at which the items were sold.
- Each dimension may have a table associated with it, called a dimension table, which further describes the dimension.
- Multidimensional databases help to provide data-related answers to complex business queries quickly and accurately. Data warehouses and Online Analytical Processing (OLAP) tools are based on a multidimensional data model. OLAP in data warehousing enables users to view data from different angles and dimensions.

Product \ Year	1999	2000	2001	2002	ALL
chairs	25	37	89	21	172
tables	10	30	0	45	85
desks	56	84	9	35	184
shelves	19	20	0	71	110
boards	5	16	11	15	47
ALL	115	187	109	187	598

Fig. 2.9.1 Data cube

- The data cube summarizes the measure with respect to a set of n dimensions and provides summarizations for all subsets of them. Fig. 2.9.1 shows data cube.
- The most detailed part of the cube is called a base cuboid. The top most 0-D cuboid, which holds the highest-level of summarization, is called the apex cuboid. The lattice of cuboids forms a data cube.

Product	Year 1999	2000	2001	2002	ALL
Chairs	25	37	89	21	172
Tables	10	30	0	45	85
Desks	56	84	9	35	184
Shelves	19	20	0	71	110
Boards	5	16	11	15	47
ALL	115	187	109	187	598

Base cuboid

Apex cuboid

Fig. 2.9.2 Base cuboid

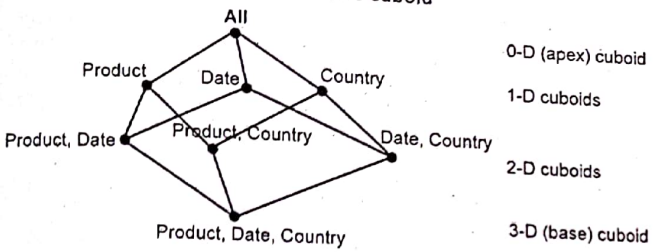


Fig. 2.9.3

- As a logical structure, a cube allows a client application to retrieve values, of measures, as if they were contained in cells in the cube; cells are defined for every possible summarized value.
- A cell, in the cube, is defined by the intersection of dimension members and contains the aggregated values of the measures at that specific intersection.

2.9.1 Stars, Snowflakes and Fact Constellations

- Star schema gives a very simple structure to store the data in the data warehouse.
- The purpose of star schema is to collect the information of numerical "fact" data relating to business and separate it from "dimensional" or descriptive data.
- Fact data includes information like weight, price, quantities and speed that is the data in the numerical format. Dimensional data includes information of untouchable things like model names, colors, employee names, geographical locations along with numerical data.
- A fact table consists of facts of a particular business process e.g., sales revenue by month by product. Facts are also known as measurements or metrics. A fact table record captures a measurement or a metric.

- A fact table is used in the dimensional model in data warehouse design. A fact table is found at the center of a star schema or snowflake schema surrounded by dimension tables.
- Dimension tables are generally used to define dimensions; they include the values, dimension keys as well as attributes. Typically, dimension tables are small in size, ranging from a few to numerous thousand rows.
- Most data warehouses use a star schema to represent multi-dimensional model. Each dimension is represented by a **dimension table** that describes it.
- A fact table connects to all dimension tables with a multiple join. Each tuple in the **fact table** consists of a pointer to each of the dimension tables that provide its multi-dimensional coordinates and stores measures for those coordinates. The links between the fact table in the center and the dimension tables in the extremities form a shape like a star.
- Fig. 2.9.4 shows star schema.

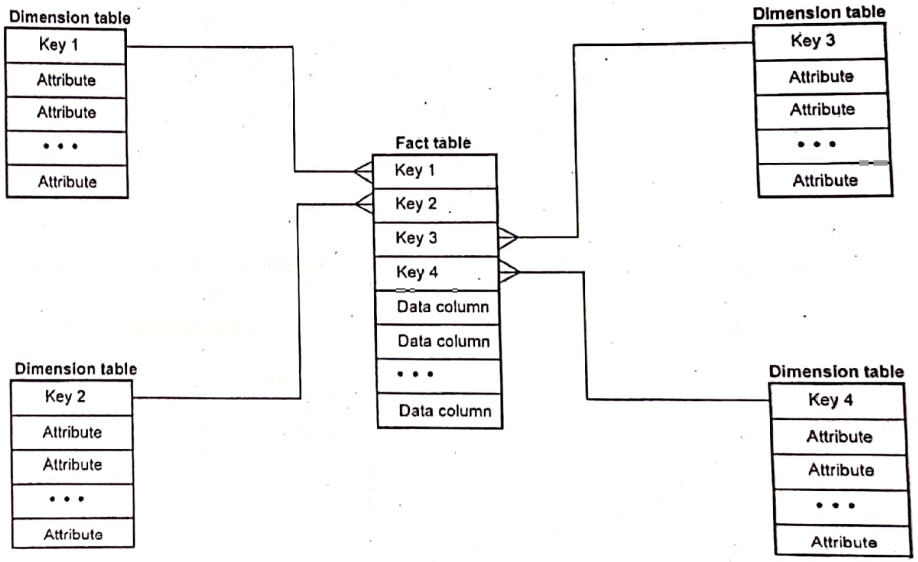


Fig. 2.9.4 Star schema

- A star schema stores all of the information about a dimension in a single table. Each level of a hierarchy is represented by a column or column set in the dimension table.

- A dimension object can be used to define the hierarchical relationship between two columns that represent two levels of a hierarchy; without a dimension object, the hierarchical relationships are defined only in metadata. Attributes are stored in columns of the dimension tables.
- A snowflake schema normalizes the dimension members by storing each level in a separate table.
- Measures are stored in fact tables. Fact tables contain a composite primary key, which is composed of several foreign keys (one for each dimension table) and a column for each measure that uses these dimensions.

Snowflakes and star schema :

- Star schema : A fact table in the middle connected to a set of dimension tables.
- Star schema consists of a fact table with a single table for each dimension.
- Snowflake schema : A refinement of star schema where some dimensional hierarchy is normalized into a set of smaller dimension tables, forming a shape similar to snowflake.
- It is a variation of star schema, in which the dimensional tables from a star schema are organized into a hierarchy by normalizing them.

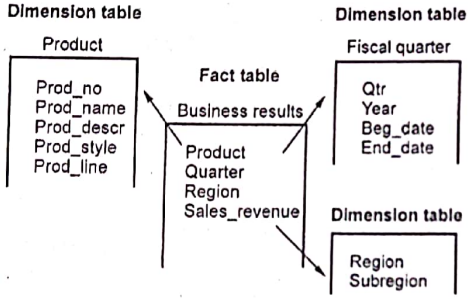


Fig. 2.9.5 A star schema with fact and dimensional tables

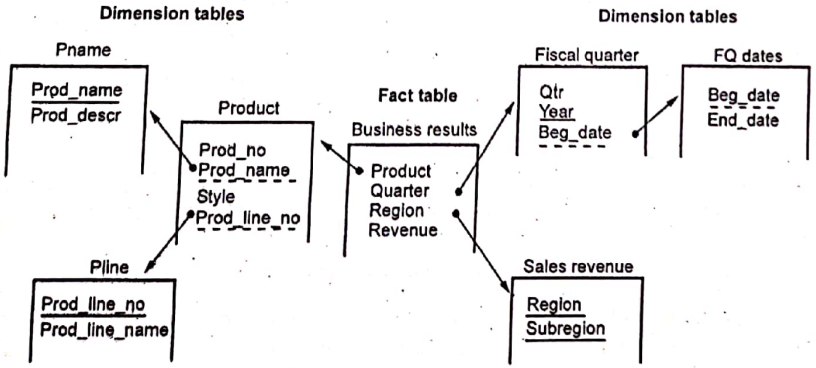


Fig. 2.9.6 Snowflake schema

- The major difference between the snowflake and star schema models is that the dimension tables of the snowflake model may be kept in normalized form to reduce redundancies. Such a table is easy to maintain and saves storage space.

2.9.2 Difference between Fact Table and Dimension Table

Parameters	Fact table	Dimension table
Basic	Fact table contains the measurement along the attributes of a dimension table.	Dimension table contains the attributes along which fact table calculates the metric.
Attributes and records	It contains less attributes and more records.	It contains more attributes and less records.
Table size	Fact table grows vertically.	Dimension table grows horizontally.
Key	Fact table contains a primary key which is a concatenation of primary keys of all dimension table.	Each dimension table contains its primary key.
Creation	Fact table can be created only when dimension tables are completed.	Dimension tables need to be created first.
Schema	Schema contains less number of fact tables.	Schema contains more number of dimension tables.
Attributes	Fact table can have data in numeric as well as textual format.	Dimension table always contains attributes in textual format.

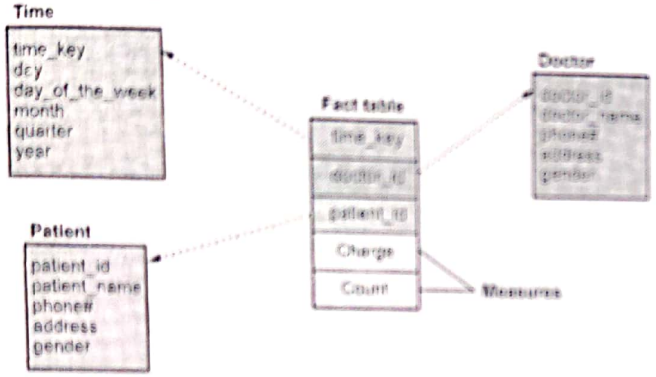
2.9.3 Compare Star Schema and Snowflake Schema

Star schema	Snowflake schema
Star schema is the simple and common modelling paradigm where the data warehouse comprises of a fact table with a single table for each dimension.	Snowflake schema is the kind of the star schema which includes the hierarchical form of dimensional tables.
Star schema does not use normalization.	This schema uses normalization to eliminate redundancy of data.
It contains fact and dimension tables.	It contains sub-dimension tables including fact and dimension tables.
Simple to understand and easily designed.	Hard to understand and design.
High level of data redundancy.	Low level of data redundancy.

Cube processing is faster.	Cube processing might be slow because of the complex join.
It uses more space.	It uses less space.

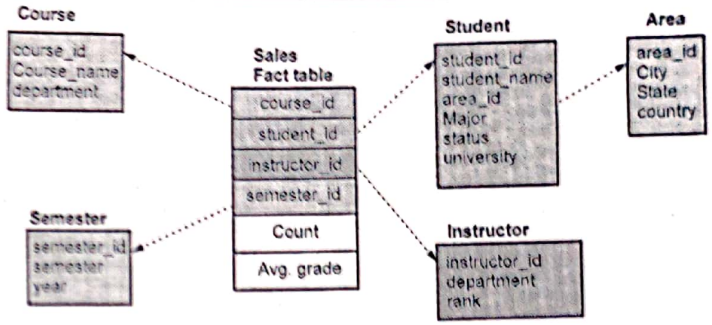
Example 2.9.1 *Suppose that a data warehouse consists of the three dimensions time, doctor and patient and the two measures count and charge, where charge is the fee that a doctor charges a patient for a visit. Draw a star's schema diagram for the data warehouse.*

Solution : Star schema diagram



Example 2.9.2 *Suppose that a data warehouse for big-university consists of the following four dimensions : Student, course, semester and instructor and two measures count and avg grade. When at the lowest conceptual level, the avg grade measure stores the actual course grade of the student. At higher conceptual levels, avg grade stores the average grade for the given combination. Draw snowflakes schema diagram.*

Solution : Snowflake's schema diagram



Review Questions

1. *Define : i) Dimensions ii) Measures iii) Fact tables.*
2. *Discuss the concept of star, snowflake and galaxy schemas for multidimensional databases.*

❑❑❑