Unit I Business Intelligence Concepts

Business Intelligence and Data Analytics

Mrs. Madhuri Prashant Karnik

Department of Computer Engineering

BRACT'S, Vishwakarma Institute of Information Technology, Pune-48

(An Autonomous Institute affiliated to Savitribai Phule Pune University)
(NBA and NAAC accredited, ISO 9001:2015 certified)

Learning Objectives

1) To learn the basics of the Business intelligence Process.

2) To understand the Decision making support system.

3) To learn the design of data warehouse.

4)To learn modeling a web based social business problem

5)To design a dashboard using visualization technique.

6)To understand different analytics techniques.

Learning Outcome/Course Outcome

1. Remember the Business intelligence concept for projects.

2. Apply Decision support system techniques for BI applications.

3. Design the data warehouse technique for business intelligence.

4. Acquire the knowledge of emerging and critical area in social media analytics.

5.Apply contemporary visualization techniques and tool for real/distinguished time applications.

6. Demonstrate analytical techniques for different case studies.

- Introduction to data
- Information and Knowledge
- Operational and Informational data
- Introduction to Business Intelligence
- BI architecture and its components
- BI opportunities, Benefits of BI
- Role of mathematical model in BI
- Factors Responsible for successful BI Project
- Obstacle to Business Intelligence in an Organization

What is Data?

- **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.
- Definition for data presented by Thierauf (1999): "unstructured facts and figures that have the least impact on the typical manager."

What is Information?

- Information is a set of data that is processed in a meaningful way according to the given requirement. It is processed, structured, or presented in a given context to make it meaningful and useful.
- Information assigns meaning and improves the reliability of the data. It helps to ensure undesirability and reduces uncertainty.
- Therefore, when the data is transformed into information, it never has any useless details. It includes data that possess context, relevance, and purpose.
- It also involves the manipulation of raw data which eventually becomes knowledge.

- **Data, information, and knowledge have significant and discrete meanings.**
- **Data are specific, objective facts or observations**
- **Information is defined as “data capable with relevance and purpose”.**

What is Knowledge?

- Knowledge is a combination of information, experience, and insight that helps the individual or the organization.
- It is linked to doing and implies know-how and understanding. Knowledge is possessed by each individual and is an outcome of his or her experience.
- It also covers the norms to evaluates new inputs from his surroundings.
- Knowledge is a mix of related information, experiences, rules, and values.
- Richer, deeper, and more valuable.

Tacit vs. Explicit Knowledge

- **Tacit knowledge is personal, context-specific and hard to formalize and communicate**
  - A [knowledge] developed and internalized by the knower over a long period of time . . .’
- **Explicit knowledge can be easily collected, organized and transferred through digital means.**
  - A theory of the world, conceived of as a set of all of the conceptual entities describing classes of objects, relationships, processes, and behavioral norms.

Tacit Knowledge

- Knowing how to identify the key issues necessary to solve a problem
- Applying similar experiences from past situations
- Estimating work required based on perception & experience
- Deciding on an appropriate course of action

Explicit Knowledge

- Procedures listed in a manual
- Books and articles
- News reports and financial statements
- Information left over from past projects

Examples of explicit and tacit knowledge

Information	Knowledge
Information is refined data	The knowledge is useful information
Data and context	Information, experience, and intuition
Comprehension of data is its outcome.	Understanding of information is its outcome.
Easily transferable	To transfer you require learning
Improves representation	Increases awareness
All information need not be knowledge.	All knowledge is information.
Information can be reproduced.	Knowledge reproduction is not possible.
Information alone is not sufficient to make any predictions.	Prediction is possible if one possesses the required knowledge.
A flow of meaningful messages	Beliefs and commitments created from these messages
A message used to change the receiver's perception	It contains experiences, values, insights, and contextual information

Information	Knowledge
Text that answers the questions a **who, when, what, or where.**	Text that answers the questions of **why and how.**
The information indicates the organized data about someone or something which is obtained from various sources like the internet, newspaper, television, etc.	Knowledge means the awareness or understanding of the subject obtained from the education or experience of a particular person.
Information is a refined form of data that is useful to understand the meaning.	knowledge is the relevant information that helps in drawing conclusions.
Processing results allow you to improve the representation and ensures an easy interpretation of the information.	Processing results in increased consciousness, therefore, enhance subject knowledge.
Information brings on comprehension of the figure and facts.	Knowledge can lead to an understanding of the subject.
The transfer of information is easy using different means. It can be verbal or non-verbal signals.	The transfer of knowledge is difficult, as it requires learning on the part of the receiver.

Examples given for data:

-4,8,12,16

-Dog, cat, cow

-161.2, 175.3, 166.4, 164.7, 169.3

Examples given for information:

-4, 8, 12 and 16 are the first four answers in the 4 x table

-Dog, cat, cow is a list of household pets

-165, 175.2, 186.3, 164.3, 169.3 are the height of 14-year old students.

Examples given for Knowledge:

-4, 8, 16 and 24 are the first four answers in the 4 x table (because the 4 x table starts at

-four and goes up in four. Similarly the 5 x table must start at five and go up in fives)

-A tiger is not a household pet as it is not in the list, and it lives in the wild forest.

-The tallest student is 186.3cm.

Operational and Informational Data

1. **Operational Systems :**

An operational system is a generally known term in data warehousing that specifies a system which is used to maintain records of daily business transactions in an organization. Operational system is also termed as Online Transaction Processing (OLTP). Operational systems have to deal with the running data values and consists of data like payroll, inventory, order ,purchase data and other daily operations on data.

2. **Informational Systems :**

Informational systems are the standardized systems that are commonly implemented within the people, processes, and technology in an organization for improving the interaction. Informational systems are designed to deals with the collection, compilation of data and deriving information from that data.

Informational systems are used everywhere for increasing the performance of the businesses and organizations. For example, financial reporting produced by a company, status reports provide critical feedback and updates on products, feedback etc.

S.NO	OPERATIONAL SYSTEMS	INFORMATIONAL SYSTEMS
1.	Operational systems are designed to deal with the running values of data.	Informational Systems deals with the collection, compilation and deriving information from data.
2.	In operational systems, optimization of data structure is done for transactions.	In informational systems, optimization of data structure is done for complex queries.
3	Operational systems are generally suited for small volumes of data.	Informational Systems are mainly designed for large volumes of data and hence convenient to use.
3.	Operational systems are process oriented.	While informational systems are subject oriented.
5.	Operational systems supports various data access operations such as read, update and delete.	Informational systems only supports read operation for data access.

Operational System(Data) and Informational system(Data)

Table 11-1 Comparison of Operational and Informational Systems

Characteristic	*Operational Systems*	*Informational Systems*
Primary purpose	Run the business on a current basis	Support managerial decision making
Type of data	Current representation of state of the business	Historical point-in-time (snapshots) and predictions
Primary users	Clerks, salespersons, administrators	Managers, business analysts, customers
Scope of usage	Narrow, planned, and simple updates and queries	Broad, ad hoc, complex queries and analysis
Design goal	Performance throughput, availability	Ease of flexible access and use
Volume	Many, constant updates and queries on one or a few table rows	Periodic batch updates and queries requiring many or all rows

Business Intelligence (BI)

- BI is an umbrella term that combines architectures, tools, databases, analytical tools, applications and methodologies.BI helps us to convert data into information, information into knowledge and knowledge into plans that guides organizations for their betterment(DSS).

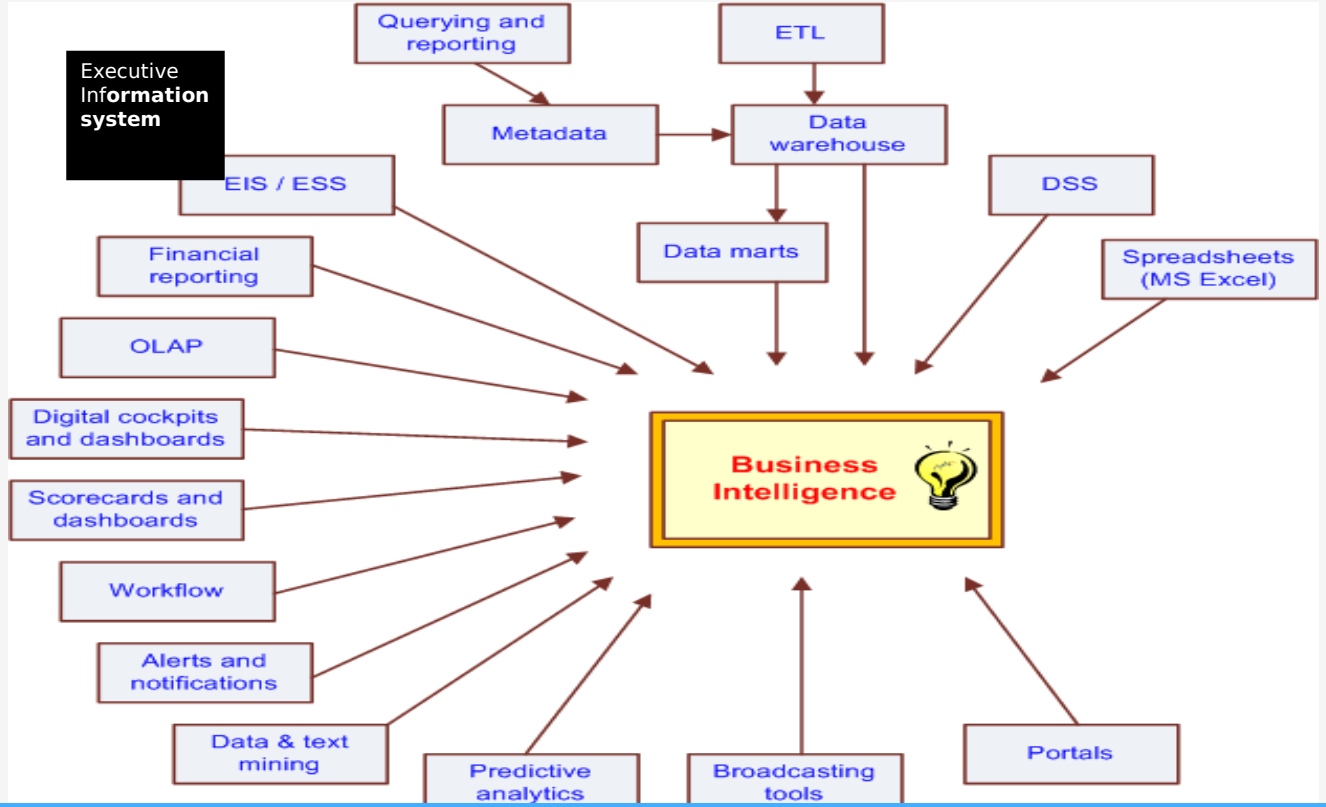
- BI helps *transform* data, to information (and knowledge), to decisions and finally to action.

- BI's major objective is to enable easy access to data (and models) to provide business managers with the ability to conduct analysis.

- The purpose of Business Intelligence is to support better business decision making. Essentially, Business Intelligence systems are data-driven Decision Support Systems (DSS).

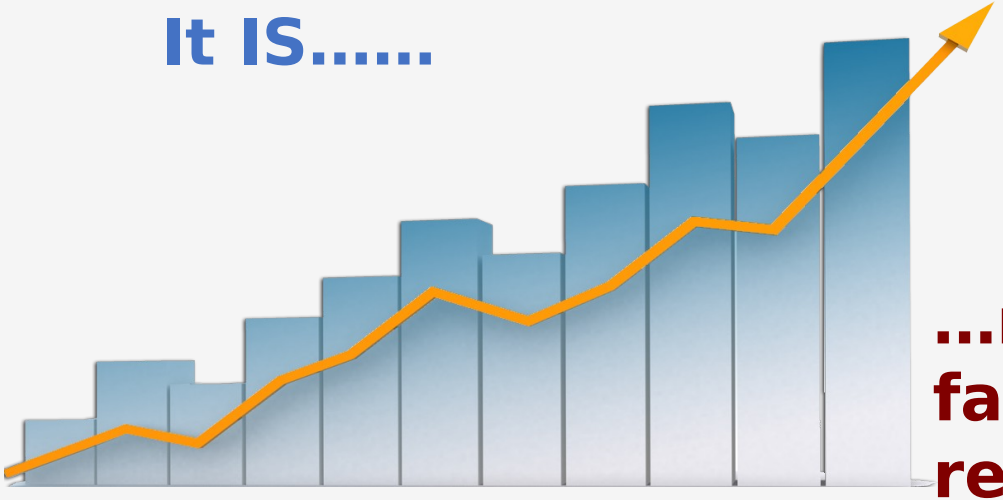
- It is a suite of software and services to transform data into actionable intelligence and knowledge.
- BI has a direct impact on organization's strategic, tactical and operational business decisions.
- BI supports fact-based decision making using historical data rather than assumptions.
- BI tools perform data analysis and create reports, summaries, dashboards, maps, graphs, and charts to provide users with detailed intelligence about the nature of the business.

The Evolution of BI Capabilities



Justify BI

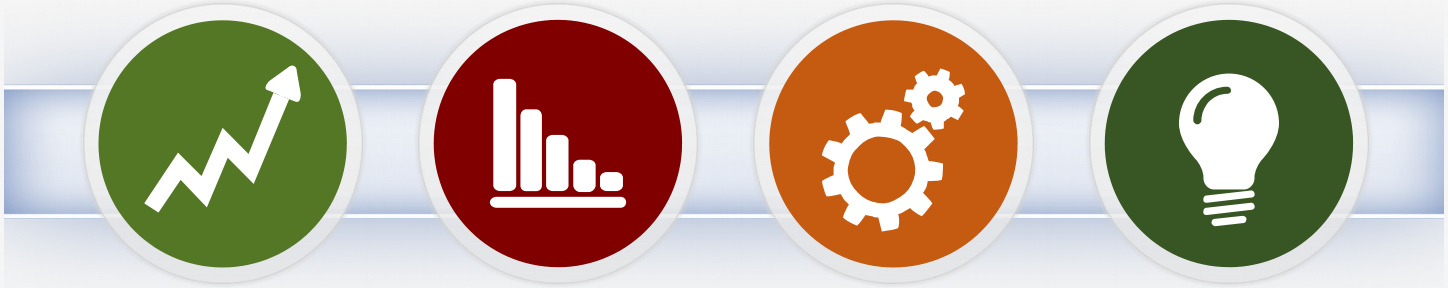
It IS……



…more than faster reporting

Justify BI

It's ALL about profitability



Increasing revenue

Reducing Costs

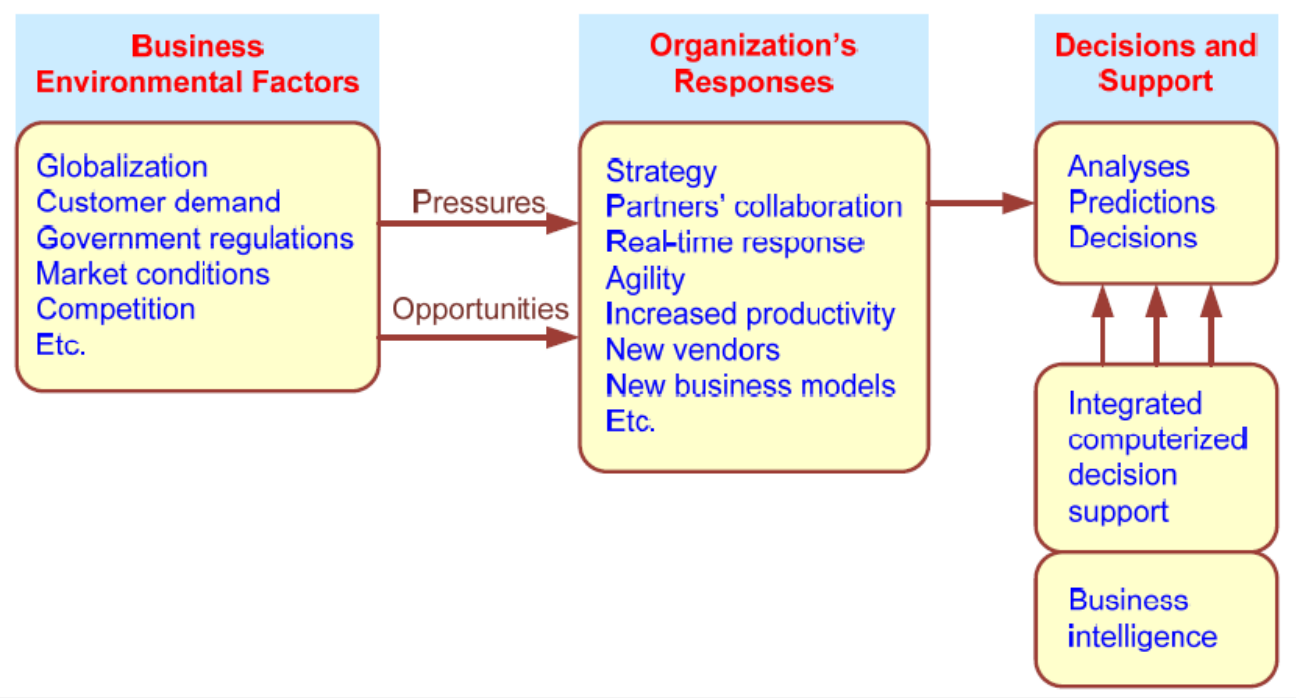
Improving Efficiencies

Better Asset Management

Changing Business Environment

- **Companies are moving aggressively to computerized support of their operations => Business Intelligence**
- **Business Pressures–Responses–Support Model**
  - Business pressures result of today's competitive business climate
  - Responses to oppose(answer) the pressures
  - Support to better facilitate the process

Business Pressures–Responses–Support Model

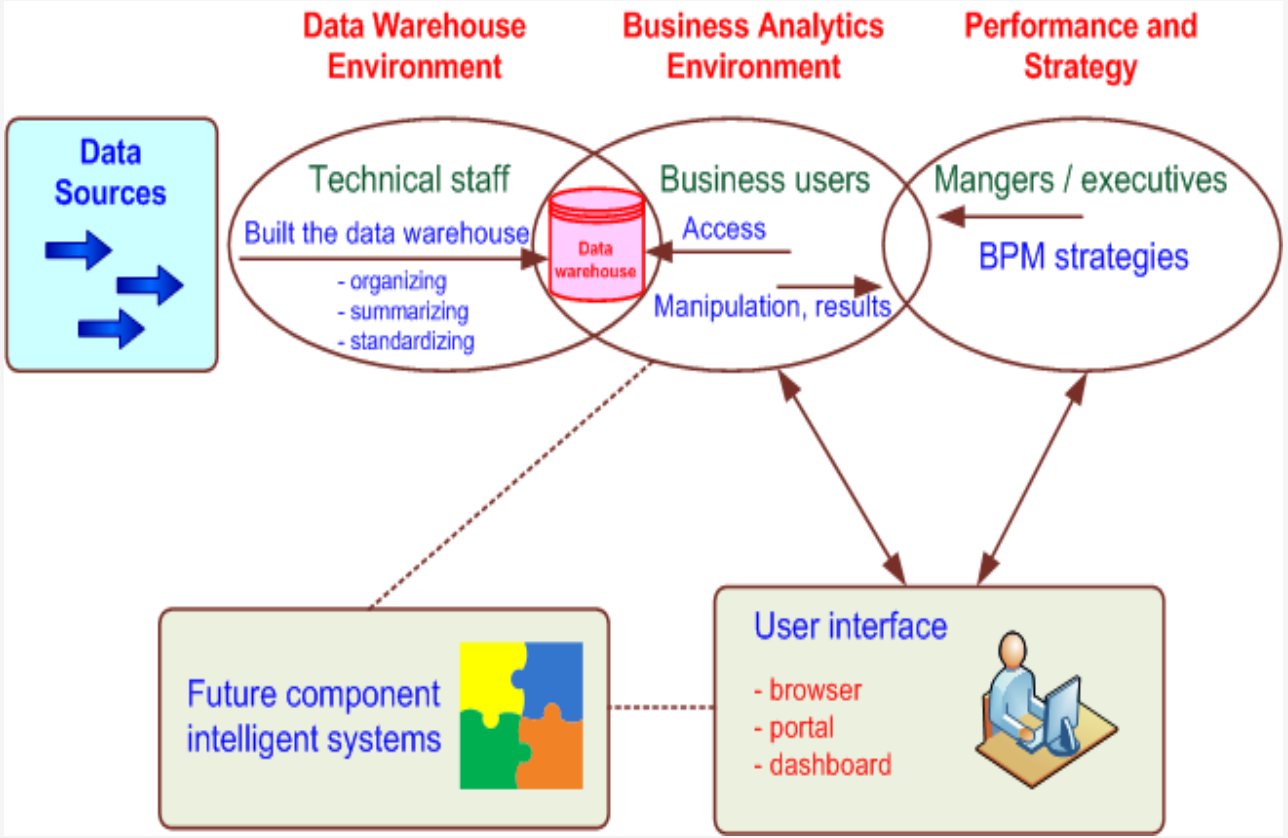


The Business Environment

- **The environment in which organizations operate today is becoming more and more complex, creating:**
  - opportunities, and
  - problems
  - Example: globalization

- **Business environment factors:**
  - markets, consumer demands, technology

A High-Level Architecture of BI



Components in a BI Architecture

- The **data warehouse** is a large repository of well-organized historical data
- **Business analytics** are the tools that allow transformation of data into information and knowledge. It is also a a collection of tools for manipulating, mining, and analyzing the data in the data warehouse.
- **Business performance management (BPM)** allows monitoring, measuring, and comparing key performance indicators by introducing the concept of management and feedback.
- **User interface** allows access and easy manipulation of other BI components. (e.g. portal, dashboard)

Example 1:

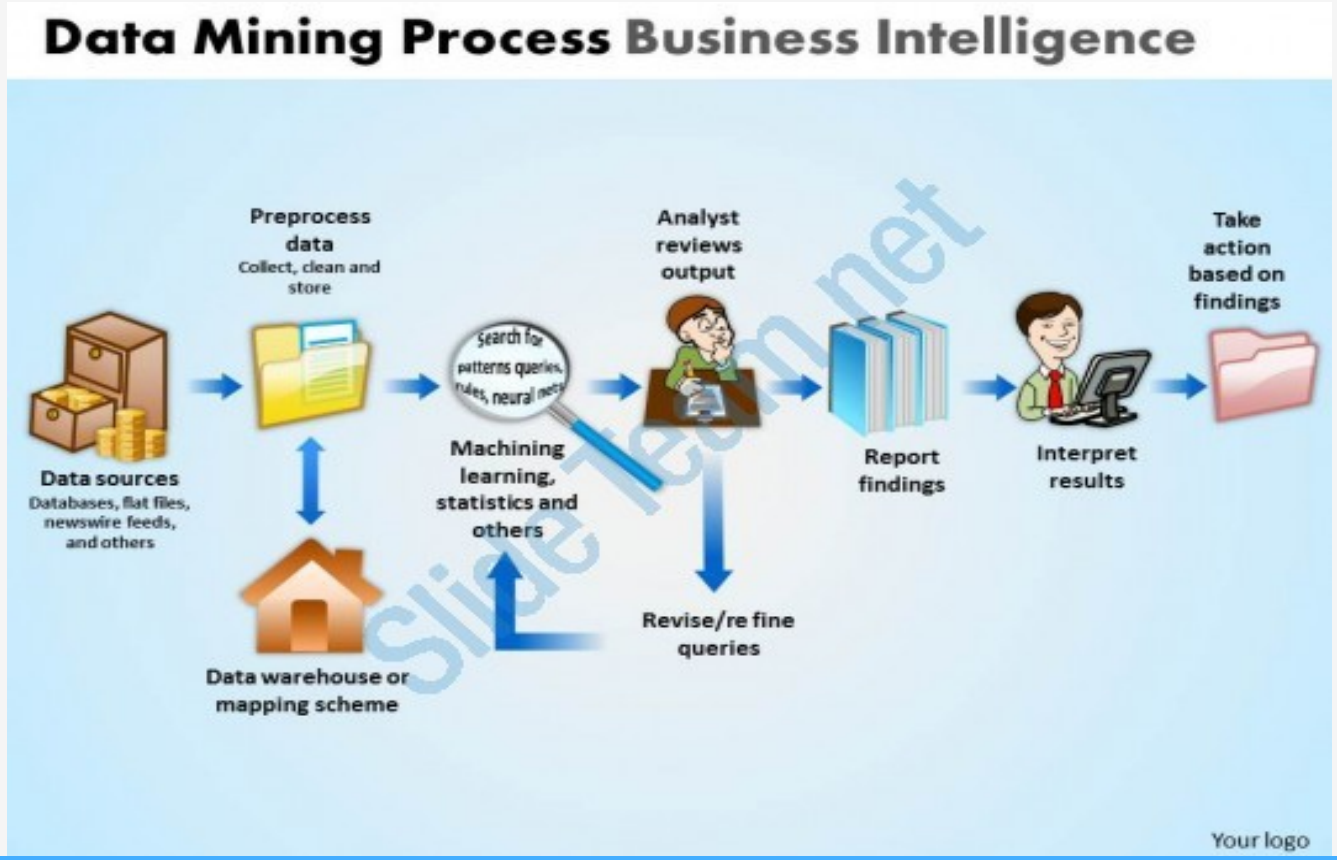
A hotel owner uses BI analytical applications to gather statistical information regarding average occupancy and room rate. It helps to find aggregate revenue generated per room.

It also collects statistics on market share and data from customer surveys from each hotel to decides its competitive position in various markets. By analyzing these trends year by year, month by month and day by day helps management to offer discounts on room rentals.

Example 2:

A bank gives branch managers access to BI applications. It helps branch manager to determine who are the most profitable customers and which customers they should work on.

BI Cycle


Data Mining Process Business Intelligence
Preprocess data
Collect, clean and store
Analyst reviews output
Take action based on findings
Search for patterns queries, rules, neural nets
Data sources
Databases, flat files, newswire feeds, and others
Machining learning, statistics and others
Report findings
Interpret results
Data warehouse or mapping scheme
Revise/re fine queries
Your logo

BI is a continuous cycle of analysis, insight, action and measurement.

Benefits of Business Intelligence

- Improve Management Processes
  - planning, controlling, measuring and/or changing resulting in increased revenues and reduced costs
- Improve Operational Processes
  - fraud detection, order processing, purchasing.. resulting in increased revenues and reduced costs
- Predict the Future-Predictive analysis.
- Better adjustment settings-competitor analysis, adjustment settings to changing trends.

Benefits of Business Intelligence

- It can eliminate a lot of the guesswork within an organization.
- Enhance communication among departments while coordinating activities
- Enable companies to respond quickly to changes in financial conditions, customer preferences, and supply chain operations.
- BI improves the overall performance of the company using it.

The DSS–BI Connection

- First, their architectures are very similar because BI evolved from DSS.
- Second, DSS directly support specific decision making, while BI provides accurate and timely information, and indirectly support decision making.
- Third, BI has an executive and strategy orientation, especially in its BPM and dashboard components, while DSS, in contrast, is oriented toward analysts.
- DSS were developed mostly in the academic world; BI were developed mostly by software companies.
- Many tools used by BI are also considered DSS tools (e.g., data mining and predictive analysis)

Factors responsible for success of the BI Project:

- Create a business case and outline the expected benefits
- Have an enterprise-wide perspective
- Establish criteria for success
- Treat information as an asset
- Adopt best practices and standards
- Set up change management procedures

- BI strategy should align with the overall IT strategy and enterprise goals
- Do a current state, future state, and gap analysis
- Think about actionable steps
- Use iterative implementation approach with parallel tracks
- Work with frameworks and adopt proven methodologies
- Assess BI readiness of the organization and identify related gaps and issues
- Document and analyse the constraints and assumptions
- Consider all BI components.

➢ Five major factors to consider in BI software

1)Platform. In all probability, your company could be using a number of different hardware / software combinations. In the future as well, your choice of specialist applications could dictate a different choice of platform to be used. Therefore, the BI solution you select must be able to support different platforms.

2)Applications. You could be running a number of different applications and some of them would be producing their own output and data. In addition, there could be large and complex applications like ERP or CRM applications running in your company.

Each of these applications will have to be studied on a case by case basis to determine the ease (or difficulty) of integrating its data with the BI solution you are planning to shortlist. In some cases, you might even have to change your applications or write bridging code to be able to integrate the solution with the application. Do remember that the lesser number of ad hoc patches you use, the more stable your final solution will be.

3)Data management. Your organization could have data in several different databases. These could be a mix of RDBMS as well as OLAP (on-line analytical processing) data. The solution you select must be able to handle data from all these sources, combine them after purification and store them in a data mart that it will then access for its queries.

4)Globalization. In case your organization has offices over several countries, the BI solution you choose must be capable of being deployed in several different languages and allow users to select from any of the languages it is capable of. Even if you do not have a presence in different countries today, you could need the solution tomorrow. Also with supply chains becoming global, it is quite possible that some of your suppliers could be in, say, Japan, in which case you may want to share some performance related data with them. At least consider having some of the major languages supported by your software.

5) Scalability. As your business grows, your needs will invariably grow as well. The solution must be able to handle a large number of users and larger datasets without requiring frequent addition of hardware. An important issue in today's environment is a move to cloud computing. Even if you haven't already done so, I anticipate that in the next two to three years, some part of your IT infrastructure would be cloud based.

Organizations generally start with storage and then move their applications and operations to the cloud as they gain more confidence. Since the move is nearly inevitable, you must anticipate this now. Check with your vendor and clarify issues about portability and what support would be forthcoming. You could even consider deploying your solution in the Internet Cloud to begin with. Doing this frees you entirely from any kind of upfront hardware investment and gives you nearly unlimited scalability.

Potential pitfalls to avoid when designing the BI strategy

- Don’t start with a narrow vision. BI strategy needs to be prepared in the context of the wider BI definition.
- Don’t plan to use any implementation approach. It has been proven that iterative implementation works better for BI initiatives.
- BI iterations should not be done in the haphazard manner. BI strategy document is the necessary roadmap that you should follow as you begin designing BI environment.
- Don’t adopt inflexible approach. BI strategy should be constantly tuned and adjusted to reflect the needs of your business.

Obstacles to BI in organization

1. The Effective Power Structure

- **In all of the public and private concerns from the junior to senior levels there is a kind of struggle for proving oneself better than others.**

- **As every employee desires to undertake the best project and prove him more proficient than others and this may result as a conflict.**

- **Business Intelligence often changes the dynamic strategies of organization and may cause a threat to the different positions and levels.**

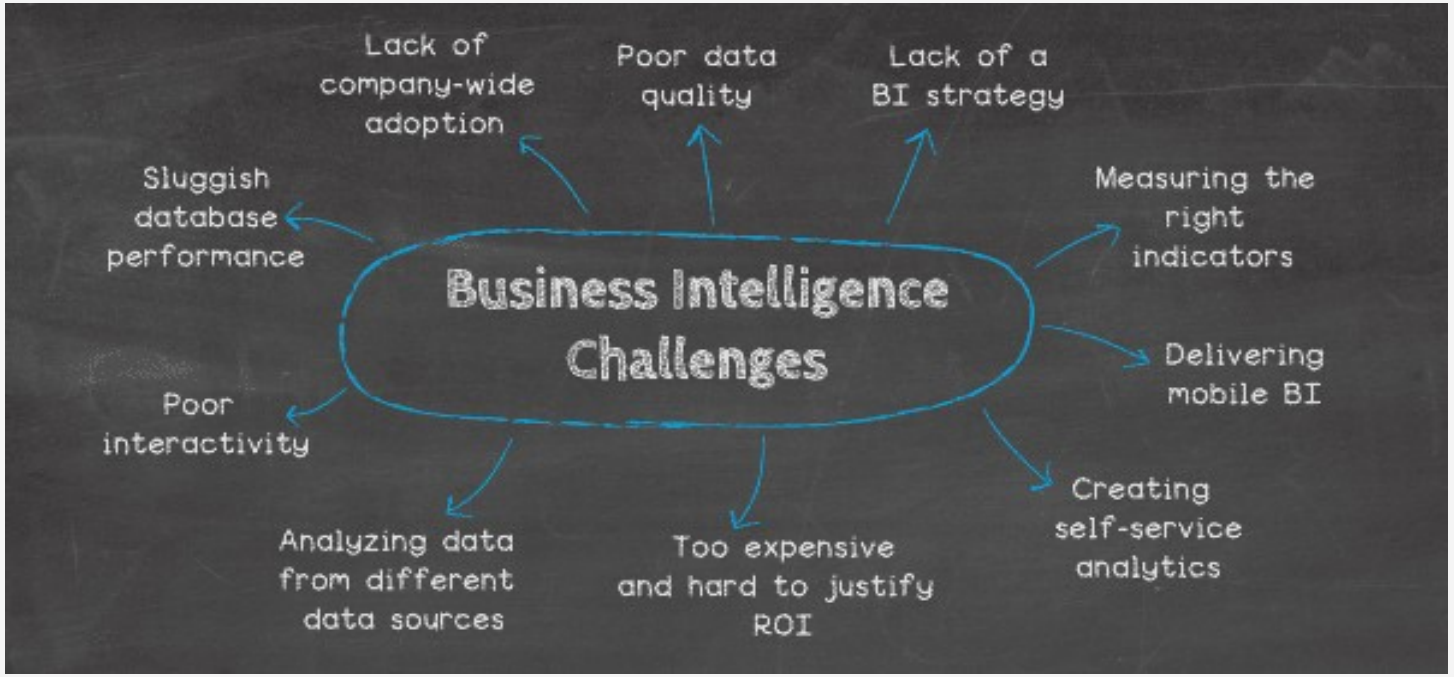
2. Recruitment and withholding of job

- **Hiring of personnel for BI applications is a responsible job.**
- **These include crisis of the suitable resource, shortage of sufficient budget, restricted schedule.**
- **Not only the recruitment as well as the retention plays a major role for the fruitful implication of BI solutions.**
- **The employees often leave the job due to disappointment to the higher reporting authority.**
- **This turns out to be as an unfavourable factor for administration of effective BI solution.**

4. Skills and Proficiency

- **The proven fact for the fulfilment of any particular business intelligence strategy considerably depends on the skills and proficiency of an employee.**

- **Due to lack of manpower resources many times a skilful responsibility is given to someone who is definitely not suitable for it.**


Business Intelligence Challenges
Lack of company-wide adoption
Poor data quality
Lack of a BI strategy
Sluggish database performance
Measuring the right indicators
Delivering mobile BI
Poor interactivity
Creating self-service analytics
Analyzing data from different data sources
Too expensive and hard to justify ROI

References

- Copyright © 2011 Pearson Education, Inc. Publishing as Prentice Hall
- https://www.comparebusinessproducts.com/business-intelligence/five-factors-to-consider-in-businessintelligence-software
- https://www.guru99.com/business-intelligence-definition-example.html
- https://www.datapine.com/blog/business-intelligence-challenges/

Text Books

- 1. R. Sharda, D. Delen, and E. Turban, Business Intelligence and Analytics. Systems for Decision Support,10th Edition. Pearson/Prentice Hall, 2015. ISBN-13: 978-0-13-305090-5, ISBN-10: 0-13- 305090-4; 3.

- 2. Introduction to business Intelligence and data warehousing, IBM, PHI, ISBN: 9788120339279

Thank You

Unit II Decision Making and Support System

Business Intelligence and Data Analytics

Mrs. Madhuri Prashant Karnik

Department of Computer Engineering

BRACT'S, Vishwakarma Institute of Information Technology, Pune-48

(An Autonomous Institute affiliated to Savitribai Phule Pune University)
(NBA and NAAC accredited, ISO 9001:2015 certified)

Unit II

Decision Making and Support System

Topics

- Concept of Decision Making system and its importance.
- Decision making process.
- Common strategies and approaches of decision makers.
- Decision support system(DSS) : Role of DSS, its main components, its various techniques.
- Types and classification.
- Applications of DSS.
- Role of Business intelligence in DSS.

- Database queries are often designed to extract specific information, such as the balance of an account or the sum of a customer's account balances.

- However, queries designed to help formulate a corporate strategy usually requires access to large amounts of data originating at multiple sources.

- A data warehouse is a repository of data gathered from multiple sources and stored under a common, unified database schema.

- Data stored in warehouse are analyzed by a variety of complex aggregations and statistical analysis.

- Furthermore, knowledge-discovery techniques may be used to attempt to discover rules and patterns from the data.

- For example, a retailer may discover that certain products tend to be purchased together, and may use that information to develop marketing strategies.

Decision-Support Systems

- Database applications can be broadly classified into transaction-processing and decision-support systems.

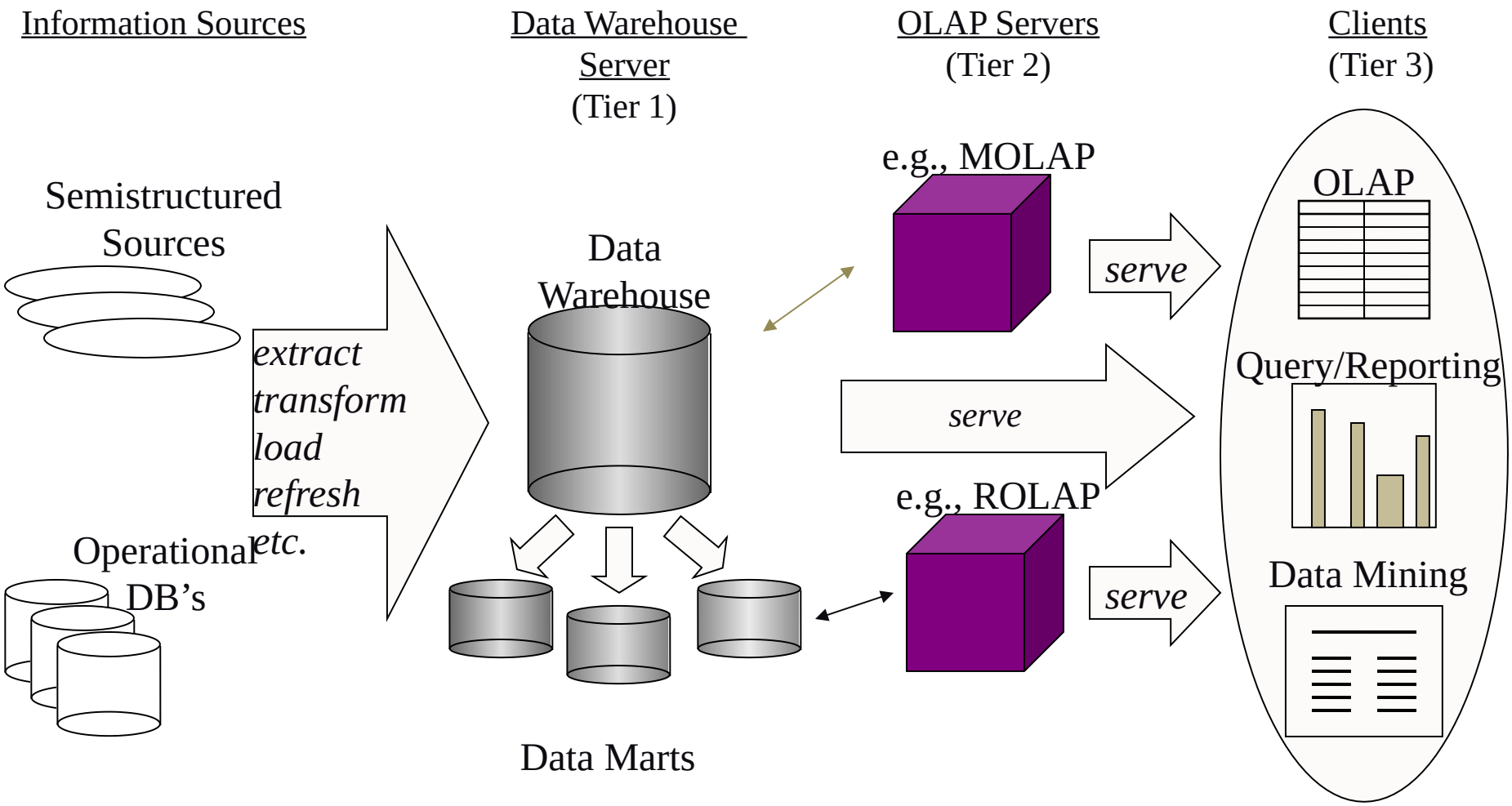
- Transaction-processing systems are systems that record information about transactions, such as product sales information for companies, or course registration and grade information for universities.

- Transaction processing systems are widely used today, and organizations have accumulated a vast amount of information generated by these systems.

•Decision-support systems aim to get high-level information out of the detailed information stored in transaction-processing systems, and to use the high-level information to make a variety of decisions.

• Decision-support systems help managers to decide what products to stock in a shop, what products to manufacture in a factory.

The Complete Decision Support System



Decision-Support Systems

- A decision support system (DSS) is a computer program application that analyzes business data and presents it so that users can make business decisions more easily.

- It is an "informational application" (to distinguish it from an "operational application" that collects the data in the course of normal business operation).

Concept of Decision Support Systems

Classical Definitions of DSS

Interactive computer-based systems, which help decision makers utilize data and models to solve unstructured problems" - *Gorry and Scott-Morton, 1971*

Decision support systems couple the intellectual resources of individuals with the capabilities of the computer to improve the quality of decisions. It is a computer-based support system for management decision makers who deal with semistructured problems - *Keen and Scott-Morton, 1978*

Structured Problems

- Structured problems are repetitive and routine problems for which standard solutions exist.

- Ex: finding an appropriate inventory level(current amount of stock), finding an optimal investment strategy.

- Management Information System(that provides managers with the tools to organize, evaluate and efficiently manage departments within an organization) primarily analyzes structured problems.

Ex. DefaulterList, Subject choice , Lecture Planning,
implementation Etc.

Semi-structured problems

- Semi-structured problems fall between structured and unstructured problems.
- Only some of the phases are structured in semi-structured problems.
- It requires a combination of standard procedures and individual judgment.
- Ex: annual evaluation of employees, setting marketing budgets for consumer products.

Unstructured problems

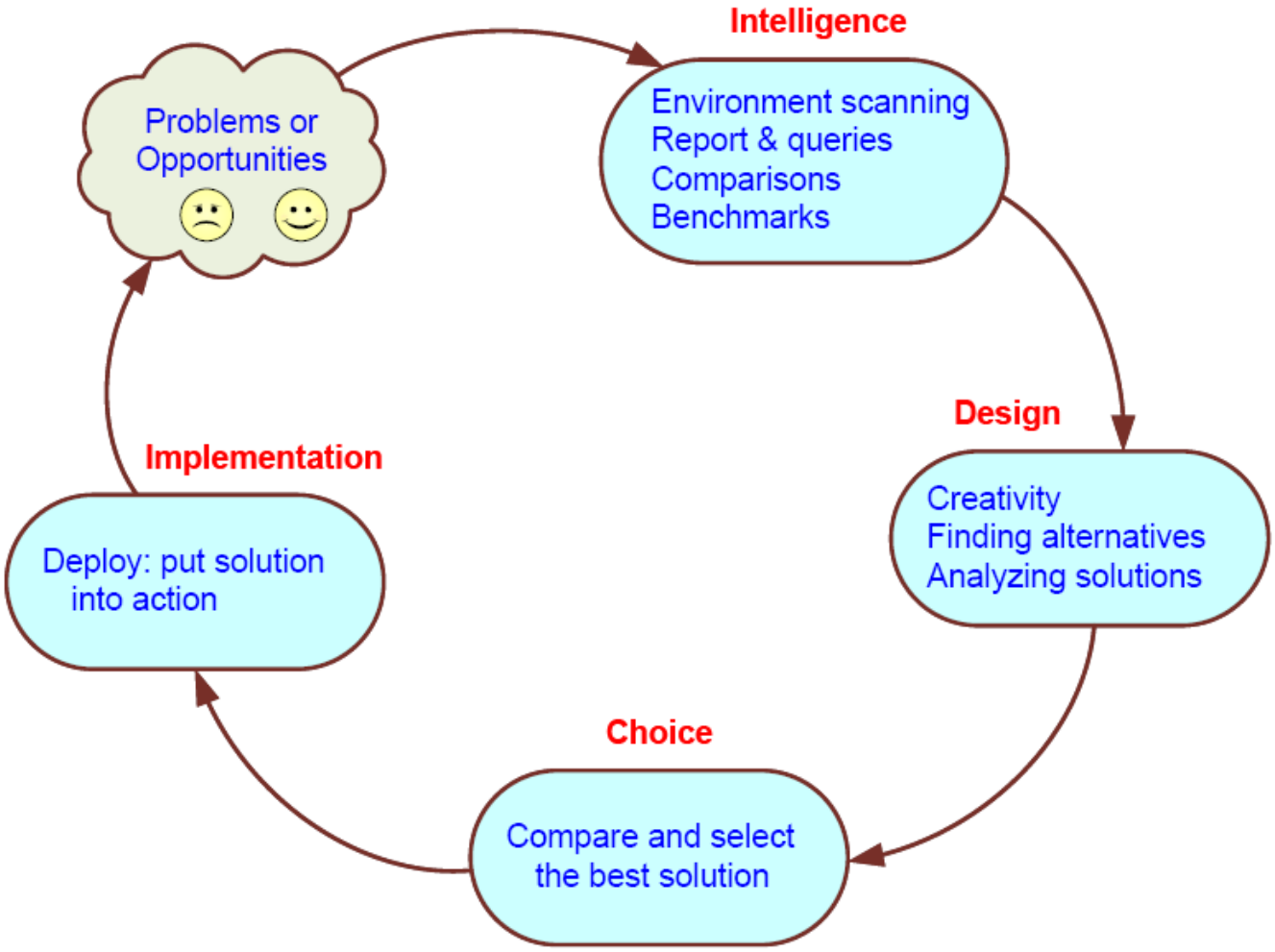
- **Unstructured problems are novel and non-routine, complex.**
- **For unstructured problems we cannot specify some procedures to make a decision.**
- **Ex: expanding the business, moving operations to foreign countries.**

A Decision Support Framework

(by Gory and Scott-Morten, 1971)

	Type of Control		
Type of Decision	**Operational Control**	**Managerial Control**	**Strategic Planning**
Structured	1 Accounts receivable Accounts payable Order entry	2 Budget analysis Short-term forecasting Personnel reports Make-or-buy	3 Financial management Investment portfolio Warehouse location Distribution systems
Semistructured	4 Production scheduling Inventory control	5 Credit evaluation Budget preparation Plant layout Project scheduling Reward system design Inventory categorization	6 Building a new plant Mergers & acquisitions New product planning Compensation planning Quality assurance HR policies Inventory planning
Unstructured	7 Buying software Approving loans Operating a help desk Selecting a cover for a magazine	8 Negotiating Recruiting an executive Buying hardware Lobbying	9 R & D planning New tech. development Social responsibility planning

Simon's Decision-Making Process



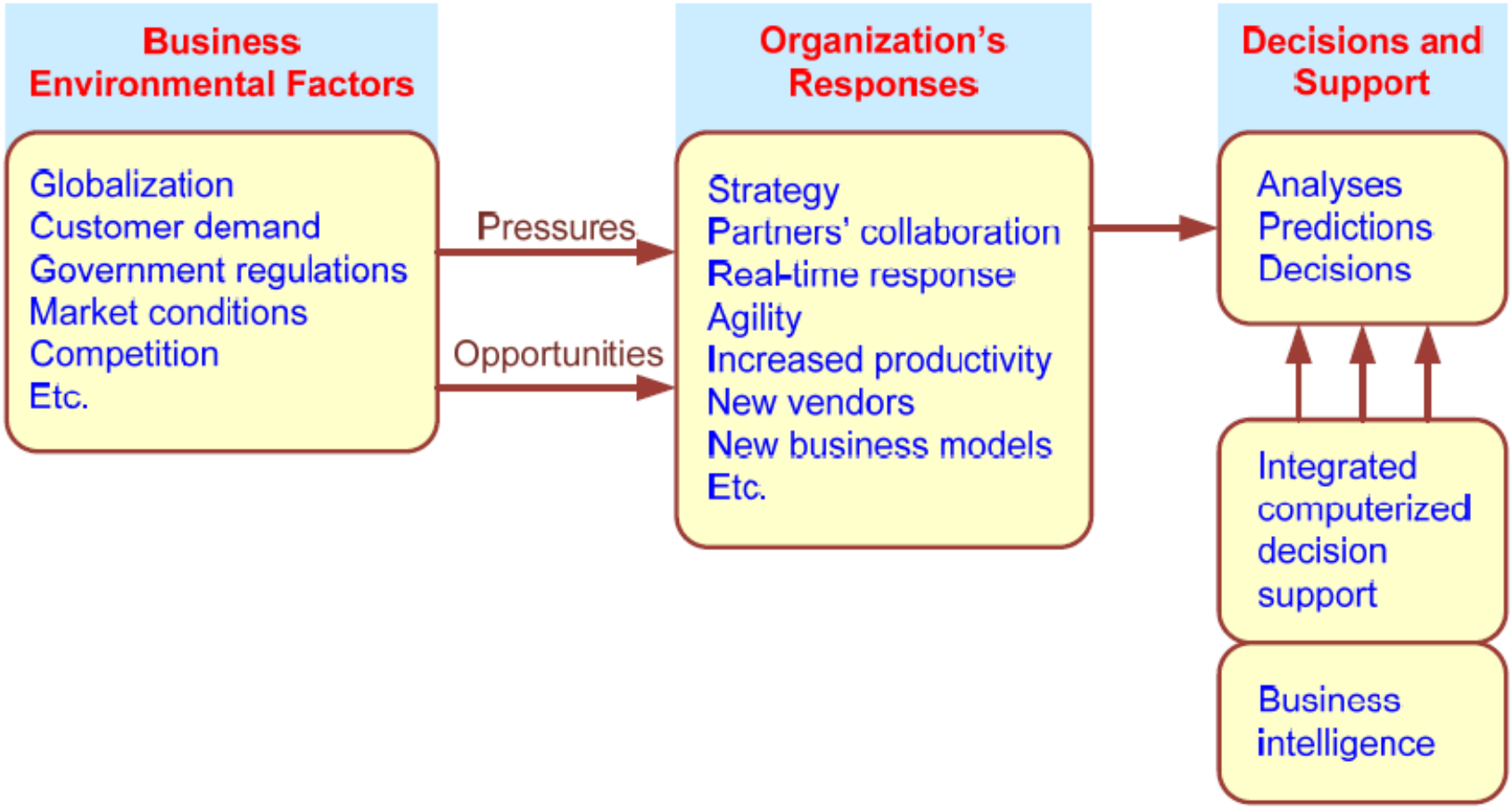
Why Decision Support System?

Changing Business Environment

- **Companies are moving aggressively to computerized support of their operations => Business Intelligence**

- **Business Pressures–Responses–Support Model**
  - Business pressures result of today's competitive business climate
  - Responses to oppose(answer) the pressures
  - Support to better facilitate the process

Business Pressures–Responses–Support Model



The Business Environment

- **The environment in which organizations operate today is becoming more and more complex, creating:**
  - opportunities, and
  - problems
  - Example: globalization

- **Business environment factors:**
  - markets, consumer demands, technology, and societal…

Organizational Responses

- **Be Reactive, Anticipate(Expressive), Adaptive, and practical**

- **Managers may take actions, such as**
  - Employ calculated planning
  - Use new and innovative business models
  - Restructure business processes
  - Participate in business alliances(agreement between multiple businesses)
  - Improve corporate information systems
  - Improve partnership relationships
  - Encourage innovation and creativity

Managers actions, continued

- Modify customer service and relationships
- Move to electronic commerce (e-commerce)
- Move to on-demand manufacturing and services
- Use new IT to improve communication, data access (discovery of information), and modify accordingly
- Respond quickly to competitors' actions (e.g., in pricing, promotions, new products and services)
- Automate certain decision processes
- Improve decision making by employing analytic

Closing the Strategy Gap by Decision Support

- One of the major objectives of computerized decision support is to facilitate closing the gap between the **current performance of an organization and its desired performance**, as expressed in its mission, objectives, and goals, and the strategy to achieve them.

Managerial Decision Making

- **Management is a process by which organizational goals are achieved by using resources**
  - Inputs: resources
  - Output: attainment of goals
  - Measure of success: outputs / inputs
- **Management ≅ Decision Making**
- **Decision making: selecting the best solution from two or more alternatives**

Decision Making Process

- **Managers usually make decisions by following a four-step process**
  1. Define the problem (or opportunity)
  2. Construct a model that describes the real-world problem
  3. Identify possible solutions to the modeled problem and evaluate the solutions
  4. Compare, choose, and recommend a potential solution to the problem

Example Detention, For BE Placement attendance , For TE internship attendance etc.

Decision making is difficult, because

- Technology, information systems, advanced search engines, and globalization result in **more and more alternatives** from which to choose.

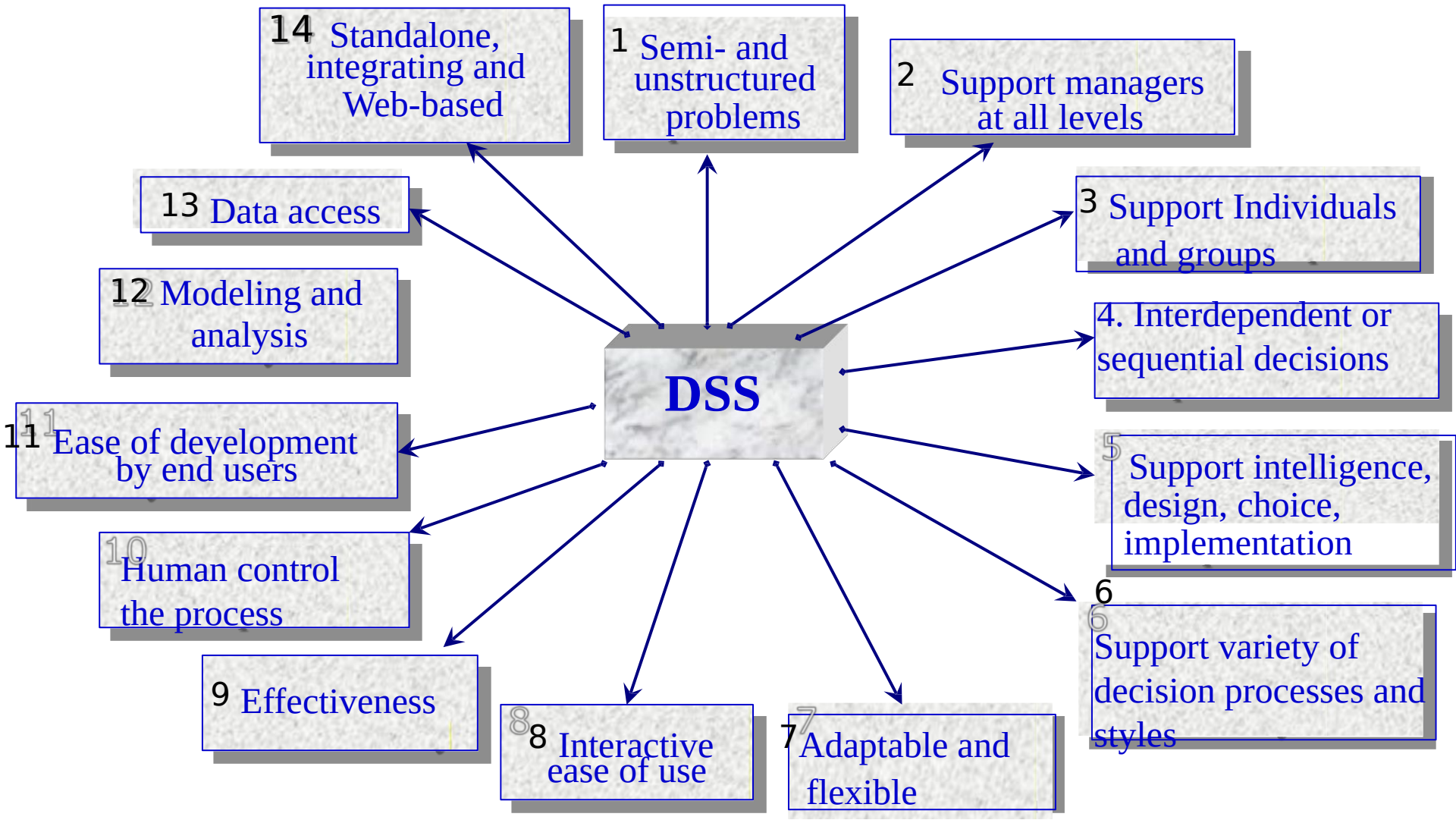
- Government regulations and the need for agreement, competition, and changing consumer demands produce more **uncertainty, making it more difficult to predict consequences and the future.**

- Other factors are the need to make rapid decisions, the frequent and unpredictable changes that make trial-and-error learning difficult, and **the potential costs of making mistakes.**

Why Use Computerized DSS

- **Computerized DSS can facilitate decision via:**
  - Speedy computations
  - Improved communication and collaboration
  - Increased productivity of group members
  - Improved data management
  - Quality support; agility support
  - Web: anywhere and anytime support

Decision Support System Description
Characteristics and Capabilities



Characteristics and Capabilities

1.Support for decision makers (mainly in semi- and un-structured situation) by bringing together human judgment and computerized information.

2.Support for all managerial levels, ranging from top executives to line managers.

3.Support for individuals (from different departments, organizational levels or different organizations) as well as groups of decision makers working somewhat independently – virtual teams through collaborative Web tools.

4.Support for independent or sequential decisions that may be made once, several times or repeatedly.

5.Support in all phases of decision-making process (*intelligence*, *design*, *choice*, *implementation*).

6.Support for a variety of decision-making process and style.

7.The decision maker should be reactive, able to tackle changing conditions quickly and able to adapt the DSS to meet these changes. DSS are flexible, so users can *add*, *delete*, *combine*, *change or rearrange basic elements*.

Characteristics and Capabilities

8. User-friendliness, strong graphical capabilities and natural language interactive human-machine interface can greatly increase the effectiveness of DSS, Most new DSS application use Web-based interfaces.
9. Improvement the effectiveness of decision making rather than its efficiency. When DSS are deployed, decision making often takes longer but the decisions are better.
10. The decision maker has complete control over all steps of the decision-making process in solving a problem – a DSS aims to support not to replace the decision maker.
11. End users are able to develop and modify simple systems by themselves. Larger systems can be built with assistance from information system specialist. Online analytical process (OLAP) and data mining software, with data warehouses, allow users to build very large and complex DSS.
12. Models are generally utilized to analyze decision-making situations. The modeling capability enable experimentation with different strategies under different configurations.

Characteristics and Capabilities

13. Access is provided to a variety of data sources, formats and types, including GIS, multimedia and object oriented.
14. Can be employed as a standalone tool used by an individual decision maker in one location or distributed throughout an organization and in several organizations along the supply chain. It can be integrated with other DSS or applications and it can be distributed internally and externally using networking and Web technologies.

These key DSS *Characteristics and Capabilities* allow decision makers to make better, more consistent decision in a timely manner and they are provided by the major DSS components.

DSS as an Umbrella Term

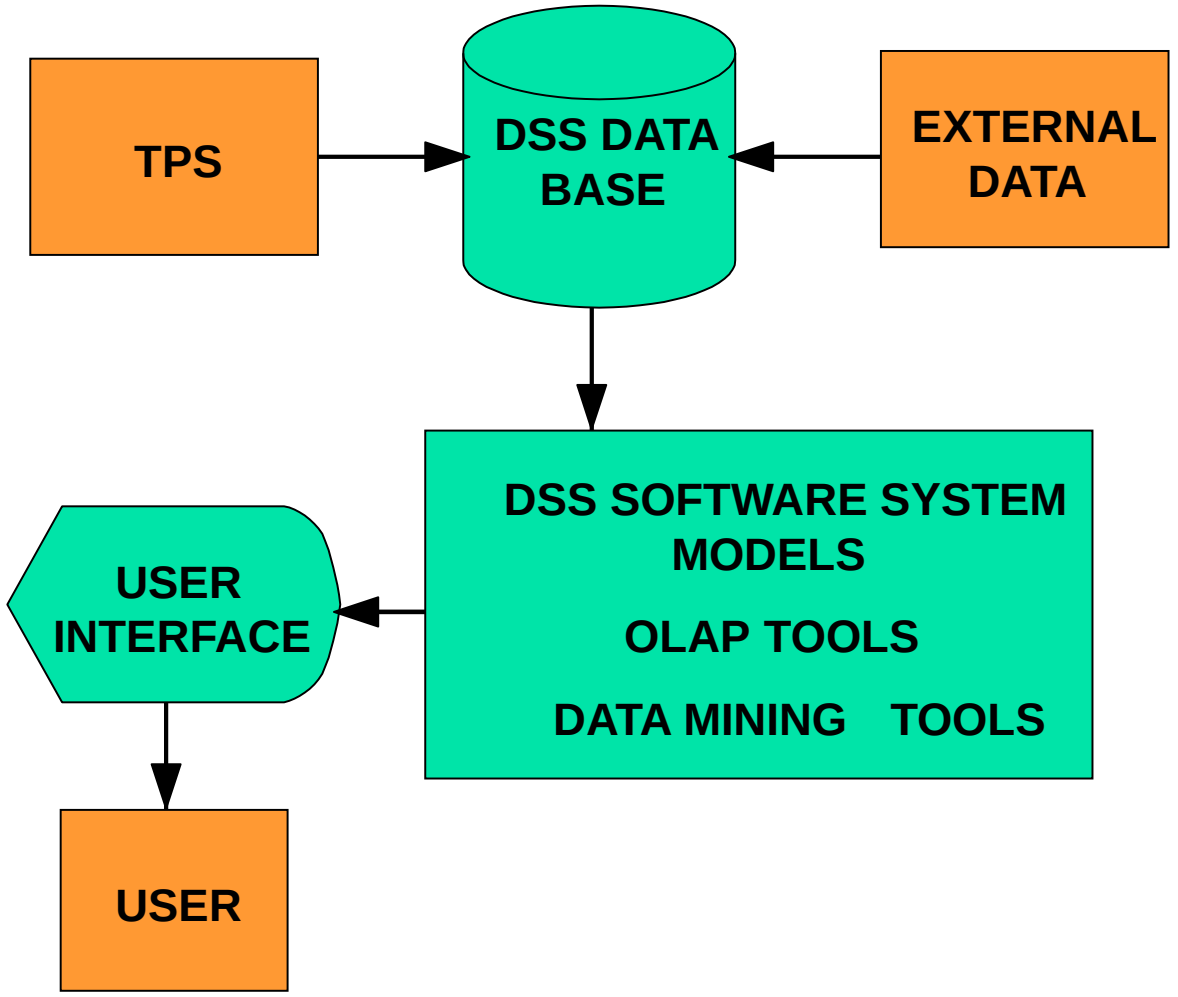
- The term DSS can be used as an umbrella term to describe **any computerized system that supports decision making** in an organization
  - E.g. A decision support system specific to an organizational function (marketing, finance, accounting, manufacturing, planning, etc.)

DSS as a Specific Application

- **In a narrow sense DSS refers to a process for building customized applications for unstructured or semi-structured problems**
- **Components of the DSS Architecture**
  - Data, Model, Knowledge/Intelligence, User Interface (API and/or user interface)
  - DSS often is created by putting together loosely coupled instances of these components

Typical Architecture

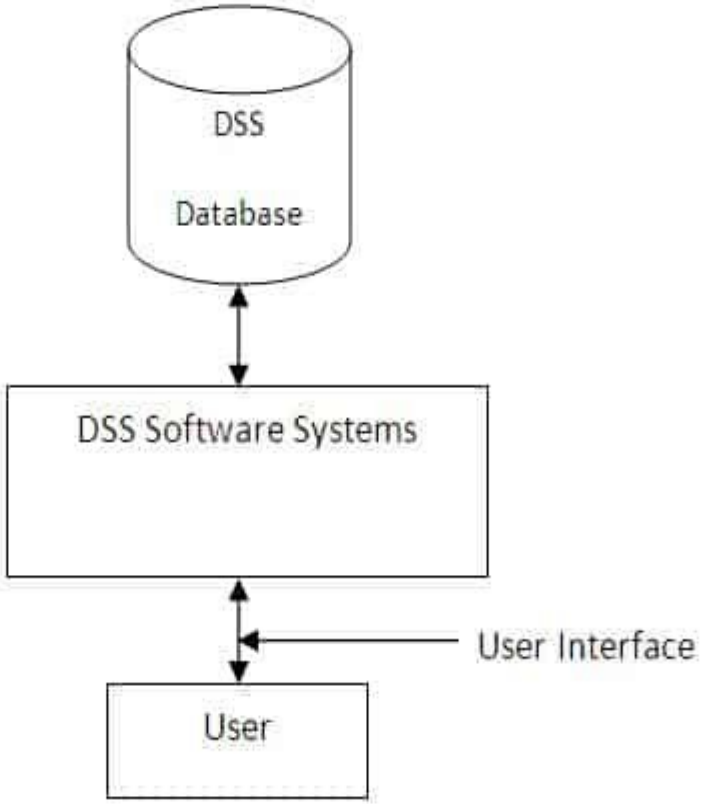
- TPS: transaction processing system
- MODEL: representation of a problem
- OLAP: on-line analytical processing
- USER INTERFACE: how user enters problem & receives answers
- DSS DATABASE: current data from applications or groups
- DATA MINING: technology for finding relationships in large data bases for prediction



Perceived benefits of DSS

- Decision Quality
- Improved Communication
- Cost Reduction
- Increased Productivity
- Time Savings
- Improved customer and employee satisfaction

Components of Decision Support Systems (DSS)



Components of Decision Support Systems (DSS)

A decision support systems consists of three main components, namely database, software system and user interface.

1. **DSS Database:** It contains data from various sources, including internal data from the organization, the data generated by different applications, and the external data mined form the Internet, etc.

The decision support systems database can be a small database or a standalone system or a huge data warehouse supporting the information needs of an organization.

Components of Decision Support Systems (DSS)

2. 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 output in the basis of different inputs 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. The selection of models that must be included in a decision support system family depends on user requirements and the purposes of DSS.

Components of Decision Support Systems (DSS)

Some of the commonly used mathematical and statistical models are as follows:-

Statistical Models: They contain a wide range of statistical functions, such as mean, median, mode, deviations etc. These models are used to establish, relationships between the occurrences of an event and various factors related to that event.

It can, for example, relate sale of product to differences in area, income, season, or other factors. In addition to statistical functions, they contain software that can analyze series of data to project future outcomes.

Components of Decision Support Systems (DSS)

Sensitivity Analysis Models: These are used to provide answers to what-if situations occurring frequently in an organization. During the analysis, the value of one variable is changed repeatedly and resulting changes on other variables are observed.

The sale of product, for example, is affected by different factors such as price, expenses on advertisements, number of sales staff, productions etc. Using a sensitivity model, price of the product can be changed (increased or decreased) repeatedly to ascertain the sensitivity of different factors and their effect on sales volume. Excel spreadsheets and Lotus 1-2-3 are often used for making such analysis.

Components of Decision Support Systems (DSS)

Optimization Analysis Models: They are used to find optimum value for a target variable under given circumstances. They are widely used for making decisions related to optimum utilization of resources in an organization. During optimization analysis, the values for one or more variables are changed repeatedly keeping in mind the specific constraints, until the best values for target variable are found.

They can, for example, determine the highest level of production that can be achieved by varying job assignments to workers, keeping in mind that some workers are skilled and their job assignment cannot be changed. Linear programming techniques and Solver tool in Microsoft excel are mostly used for making such analysis.

Components of Decision Support Systems (DSS)

Forecasting Models: They use various forecasting tools and techniques, including the regression models, time series analysis, and market research methods etc., to make statements about the future or to predict something in advance.

They provide information that helps in analyzing the business conditions and making future plans. These systems are widely used for forecasting sales.

Components of Decision Support Systems (DSS)

Backward Analysis Sensitivity Models: Also known as goal seeking analysis, the technique followed in these models is just opposite to the technique applied in sensitivity analysis models. In place of changing the value of variable repeatedly to see how it affects other variables, goal seeking analysis sets a target value for a variable and then repeatedly changes other variables until the target value is achieved.

To increase the production level by 40 percent using the backward sensitivity analysis, for example, first, the target value for the production level can be set and then the required changes to made in other factors, such as the amount of raw material, machinery and tools, number of production staff, etc., to achieve the target production level.

Components of Decision Support Systems (DSS)

3. DSS User Interface: It is an interactive graphical interface which makes the interaction easier between the DSS and its users. It displays the results (output) of the analysis in various forms, such as text, table, charts or graphics. The user can select the appropriate option to view the output according to his requirement.

A manager, for example, would like to view comparative sales data in tabular form whereas an architect creating a design plan would be more interested in viewing the result of analysis in a graphical format. The present-day decision support system built using the Web-based interface provides its users some special capabilities like better interactivity, facility for customization and personalization, and more ease of use.

Exercise

https://forms.gle/dFdAXsPkBdBC4GoQ9

Types of DSS

- **Five major types:**
  - Model-Driven DSS
  - Data-Driven DSS
  - Communication Driven DSS
  - Document Driven DSS
  - Knowledge Driven DSS

Model-driven DSS

- Model driven DSSs are complex systems that help analyze decisions or choose between different options.

- A model driven DSS emphasizes access to and manipulation of financial, optimization and / or simulation models.

- Simple quantitative models provide the most elementary level of functionality.

- Model-driven DSS use limited data and parameters provided by decision makers to help in analyzing a situation, but in general large data bases are not needed for model-driven DSS.

Model-driven DSS

•These are used by managers and staff members of a business, or people who interact with the organization, for a number of purposes depending on how the model is set up.

• These DSSs can be deployed via software / hardware in stand-alone PCs, client/server systems or the web.

Model-driven DSS

- These systems are standalone systems and they are not connected with other major corporate information systems.

- The capability of analysis of these systems is supported by some strong theory (or model) along with a good user interface that makes them easy to use.

- The use of various models in these systems helps them to perform what-if and other similar analyses. They are used for creating simulation models, performing production planning and scheduling, and creating statistical and financial reports.

Model-driven DSS

- A model-driven DSS emphasizes access to and manipulation of a statistical, financial, optimization, or simulation model.
- Model-driven DSS use data and parameters provided by users to assist decision makers in analyzing a situation;
- Dicodess (A software framework for developing Distributed co-operative decision support system) is an example of an open source model-driven DSS generator.
- Other examples:
  - **A spread-sheet with formulas in**
  - **A statistical forecasting model-**statistical forecasting implies the use of statistics based on historical data to project what could happen out in the future. This can be done on any quantitative data: Stock Market results, sales, Housing sales, etc

Advantages and Disadvantages of Modeling

- Advantages
  - Less expensive than custom approaches or real(actual) systems.
  - Faster to construct than real systems
  - Less risky than real systems
  - Provides learning experience (trial and error)
  - Future projections are possible
  - Can test assumptions

- Disadvantages
  - Assumptions about reality may be incorrect
  - Accuracy of predications often unreliable
  - Requires conceptual thinking

Data-driven (retrieving) DSS

- **A data-driven DSS or data-oriented DSS emphasizes access to and manipulation of a time series of internal company data and sometimes, external data.**

- **Simple file systems accessed by query and retrieval tools provides the basic level of functionality. Data warehouses provide additional functionality. OLAP provides highest level of functionality.**

- **Examples:**
  - Accessing data from database.

Data-driven (retrieving) DSS

- Data driven DSS model puts its emphasis on collected data that is then manipulated to fit the decision maker's needs.

- Most data driven DSSs are targeted at managers, staff and also product / service suppliers.

- It is used to query a database or data warehouse to seek specific answers for specific purposes. It is deployed via a main frame system, client server link or via web.

- The main techniques that are mostly used in data-based DSS for analyzing the data are online analytical processing (OLAP) and data mining.

Communication-driven DSS

- A communication-driven DSS use network and communication technologies to facilitate collaboration on decision making.
- It supports more than one person working on a shared task.
- examples include integrated tools like Microsoft's Net Meeting or Video conferencing.
- It is related to group decision support systems.

Communication-driven DSS

•A communication driven DSS supports more than one person working on a shared task. Many collaborators work together to come up with a series of decision to set in motion a solution or strategy.

•Most communications driven DSSs are targeted at internal teams, including partners. The most commons technology used to deploy the DSS is a web or a client server.

• In general, groupware, bulletin boards, audio and video conferencing are the primary technologies for communication driven decision support.

Document-driven DSS

- A **document-driven DSS** uses storage and processing technologies to **document retrieval and analysis**.
- It manages, retrieves and manipulates unstructured information in a variety of electronic formats.
- Document database may include: Scanned documents, hypertext documents, images, sound and video.
- A **search engine** is a primary tool associated with document driven DSS.

Document-driven DSS

- Document driven DSSs are more common, targeted at a broad base of user groups.
- The purpose of such a decision support system is to search web pages and find documents on a specific set of keywords or search terms.
- This model uses computer storage and processing technologies to provide document retrieval and analysis. A document driven DSS model uses documents in a variety of data type such as text documents, spreadsheets and database records to come up with decisions and manipulate the information to refine strategies.
- The usual technology used to set up such decision support systems are via web or a client / server system.

Knowledge-driven DSS

- **A knowledge-driven DSS provides specialized problem solving expertise stored as facts, rules, procedures or in similar structures. It suggest or recommend actions to managers.**

- **Examples:**

1. MYCIN: A rule based reasoning program which help physicians diagnose blood disease.

2.Expert system

Growth of DSS into Business Intelligence

- Use of DSS moved from specialist to managers, and then whomever, whenever, wherever
- Enabling tools like OLAP, data warehousing, data mining, intelligent systems, delivered via Web technology have collectively led to the term “business intelligence” (BI) and “business analytics”

Knowledge-driven DSS

- Knowledge driven DSSs are a catch-all category covering a broad range of systems covering users within the organization setting it up, but may also include others interacting with the organization.
- It is essentially used to provide management advice or to choose products or services. Knowledge-driven DSS can suggest or recommend actions to managers. These DSS are person-computer systems with specialized problem-solving expertise.
- The expertise consists of knowledge about a particular domain, understanding of problems within that domain, and skill at solving some of these problems.
- The typical deployment technology used to set up such systems could be client / server systems, the web, or software running on stand-alone PCs

Operational System(Data) and Informational system(Data)

Table 11-1 Comparison of Operational and Informational Systems

Characteristic	*Operational Systems*	*Informational Systems*
Primary purpose	Run the business on a current basis	Support managerial decision making
Type of data	Current representation of state of the business	Historical point-in-time (snapshots) and predictions
Primary users	Clerks, salespersons, administrators	Managers, business analysts, customers
Scope of usage	Narrow, planned, and simple updates and queries	Broad, ad hoc, complex queries and analysis
Design goal	Performance throughput, availability	Ease of flexible access and use
Volume	Many, constant updates and queries on one or a few table rows	Periodic batch updates and queries requiring many or all rows

The DSS–BI Connection

- First, their architectures are very similar because BI evolved from DSS
- Second, DSS directly support specific decision making, while BI provides accurate and timely information, and indirectly support decision making(Ex. Detention)
- Third, BI has an executive and strategy orientation, especially in its BPM and dashboard components, while DSS, in contrast, is oriented toward analysts

The DSS–BI Connection – cont.

- **Fourth, most BI systems are constructed with commercially available tools and components, while DSS is often built from scratch**
- **Fifth, DSS methodologies and even some tools were developed mostly in the academic world, while BI methodologies and tools were developed mostly by software companies**
- **Sixth, many of the tools that BI uses are also considered DSS tools (e.g. data mining and predictive analysis are core tools in both)**

Applications of DSS

- **There are theoretical possibilities of building such systems in any knowledge domain.**
  - Clinical decision support system for medical diagnosis.
  - a bank loan officer verifying the credit of a loan applicant
  - an engineering firm that has bids on several projects and wants to know if they can be competitive with their costs.
  - DSS is extensively used in business and management. Executive dashboards and other business performance software allow faster decision making, identification of negative trends, and better allocation of business resources.
  - A growing area of DSS application, concepts, principles, and techniques is in agricultural production, marketing for sustainable development.
  - A specific example concerns the Canadian National Railway system, which tests its equipment on a regular basis using a decision support system.
  - A DSS can be designed to help make decisions on the stock market, or deciding which area or segment to market a product toward.

Applications

Common Day-to-Day Decision Support System Examples:

Decision support systems operate at many levels, and there are many examples in common day-to-day use.

For example,

GPS route planning determines the fastest and best route between two points by analyzing and comparing multiple possible options. Many GPS systems also include traffic avoidance capabilities that monitor traffic conditions in real time, allowing motorists to avoid congestion.

Farmers use crop-planning tools to determine the best time to plant, fertilize and earn.

Medical diagnosis software that allows medical personnel to diagnose illnesses is another example.

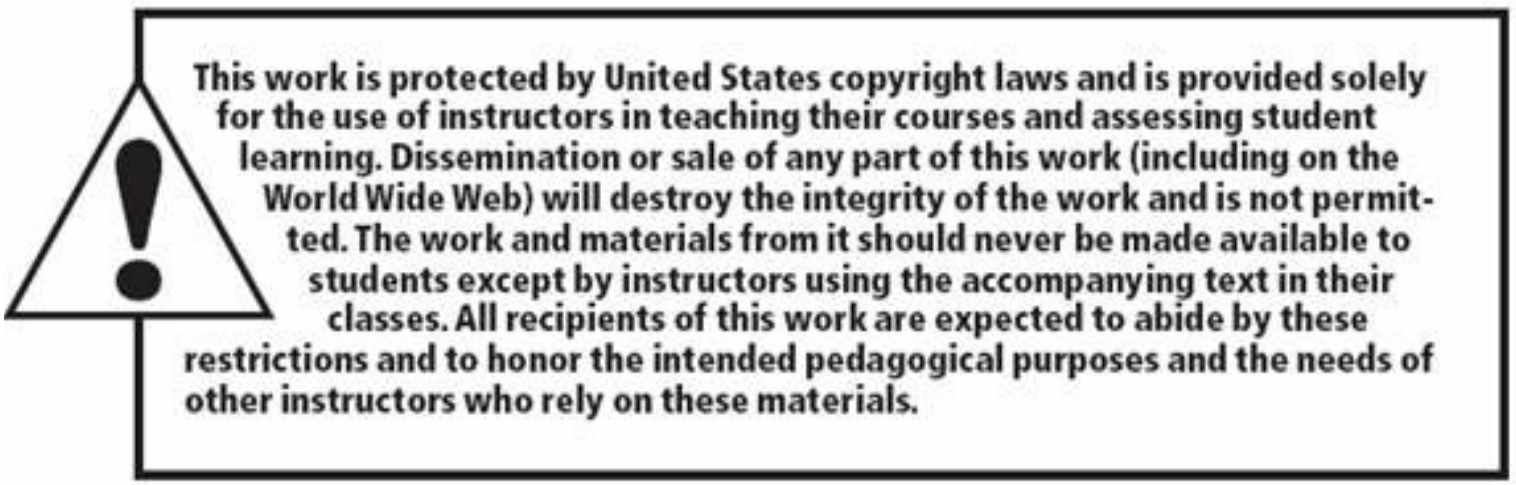
Manual and Hybrid Decision Support System Examples

Numerous manual techniques exist that support decision-making. These include activities such as the SWOT analysis where teams determine their organization's strengths and weaknesses as well as identifying threats facing the organization and potential opportunities for further growth.

The outcomes of a SWOT analysis are actionable decisions for moving the organization forward.

References

1) https://www.managementstudyhq.com/components-of-decision-support-systems.html
2) https://towardsdatascience.com/zomato-bangalore-data-analysis-6ee83652890f
3) cbafaculty.org › 23_BI › turban_dss9e_ch01



Thank You

Unit III

Data Warehouse

Unit III Data Warehouse

Business Intelligence and Data Analytics

Mrs. Madhuri Prashant Karnik
madhuri.chavan@viit.ac.in

Department of Computer Engineering

BRACT'S, Vishwakarma Institute of Information Technology, Pune-48

(An Autonomous Institute affiliated to Savitribai Phule Pune University)
(NBA and NAAC accredited, ISO 9001:2015 certified)

Topics

- Introduction
- Data Warehouse Modeling: Data Cube and OLAP Data Warehouse Design and Usage
- Distributed Data-warehouse and materialized view
- Different types of OLAP and their applications
- Difference between OLAP and OLTP
- Big Data Lakes

Data warehouse

- A **data warehouse is a repository (or archive)** of information gathered from multiple sources, stored under a unified schema, at a single site.

- Once gathered, the data are stored for a long time.

- Thus, data warehouses provide the user a single combined interface to data, making decision-support queries easier to write.

What is Data Warehouse?

- "A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management's decision-making process."—W. H. Inmon (father of data warehousing).
- A data warehouse is simply a single, complete, and consistent store of data obtained from a variety of sources and made available to end users in a way they can understand and use it in a business context
- **Data warehousing:**
  - The process of constructing and using data warehouses.

Data Warehouse—Subject-Oriented

- The Data warehouse is **subject oriented** because it provide us the information around a subject rather the organization's ongoing operations.
- These subjects can be product, customers, suppliers, sales etc.
- The data warehouse does not focus on the ongoing operations rather it focuses on modeling and analysis of data for decision making.
- Provide a simple and concise/brief view around particular subject issues by excluding data that are not useful in the decision support process.

Data Warehouse—Integrated

- Data Warehouse is constructed by integration of data from heterogeneous sources such as relational databases, on-line transaction records etc.
- This integration enhance the effective analysis of data.
- Data cleaning and data integration techniques are applied.
  - Ensure consistency in naming conventions, encoding structures, attribute measures, etc. among different data sources
  - When data is moved to the warehouse, it is converted.

Data Warehouse—Time Variant

- The Data in Data Warehouse is identified with a particular time period. The data in data warehouse provide information from historical point of view.
- The time horizon /scope for the data warehouse is significantly longer than that of operational systems
  - Operational database: current value data
  - Data warehouse data: provide information from a historical perspective (e.g., past 5-10 years)
- Every key structure in the data warehouse
  - Contains an element of time
  - But the key of operational data may or may not contain “time element”

Data Warehouse—Nonvolatile

- Non volatile means that the previous data is not removed when new data is added to it.
- The data warehouse is kept separate from the operational database therefore frequent changes in operational database is not reflected in data warehouse, that is a physically separate store of data transformed from the operational environment.
- Operational update of data does not occur in the data warehouse environment
  - Does not require transaction processing, recovery, and concurrency control mechanisms
  - Data warehouse is read only.
  - Requires only two operations in data accessing:
  - *initial loading of data* and *access of data*

Data Warehouse—Metadata

- Metadata is simply defined as data about data.
- The data that are used to represent other data is known as metadata.

- For example the index of a book serve as metadata for the contents in the book.

- It means the metadata is the summarized data that lead us to the detailed data.

Operational Data

- It is used to run the buisness.
- This data is typically stored,retrived and updated by OLTP(Online Transaction Processing) system.
- Example of OLTP: a reservation system, an accounting application, or an order-entry application.

Informational Data

- It is created from operational data.

- Informational data is typically
-summarized operational data.
-Denormalized and replicated data.
-possibly read only
-Stored on separate systems etc.

•Why Data Warehouse Separated from Operational Databases?

•The operational database is constructed for well known tasks and workload such as searching particular records, indexing etc but the data warehouse queries are often complex and it presents the general form of data.

•Operational databases supports the concurrent processing of multiple transactions. Concurrency control and recovery mechanism are required for operational databases to ensure robustness and consistency of database.

•Operational database query allow to read, modify operations while the OLAP query need only **read only** access of stored data.

•Operational database maintain the current data on the other hand data warehouse maintain the historical data.

Why Separate Data Warehouse?

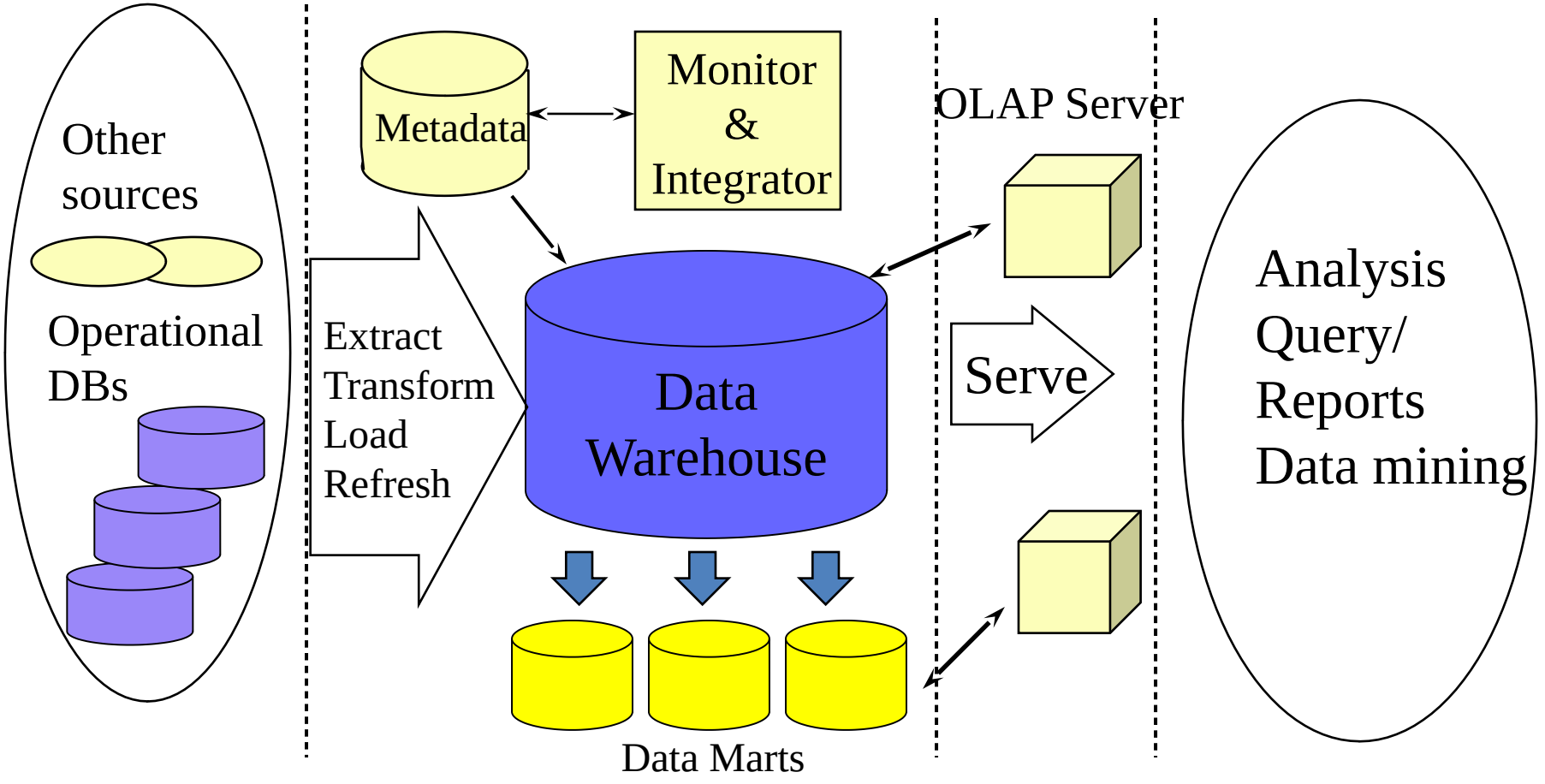
- **DBMS—** tuned for OLTP: access methods, indexing, concurrency control, recovery.

- **Warehouse—**tuned for OLAP: complex OLAP queries, multidimensional view, consolidation.

•Different functions and different data:

- missing data: Decision support requires historical data which operational DBs do not typically maintain
- data consolidation: DBs requires consolidation (aggregation, summarization) of data from heterogeneous sources
- data quality: different sources typically use inconsistent data representations, codes and formats which have to be reconcile

Data Warehouse: A Multi-Tiered Architecture



Data Warehouse Back-End Tools and Utilities

- **Data extraction**
  - get data from multiple, heterogeneous, and external sources
- **Data cleaning**
  - detect errors in the data and rectify them when possible
- **Data transformation**
  - convert data from host format to warehouse format
- **Load**
  - sort, summarize, consolidate, compute views, check integrity, build and partitions
- **Refresh**
  - propagate (transmit) the updates from the data sources to the warehouse

Three-Tier Architecture

- **Warehouse database server**
  - Almost always a relational DBMS, unstructured data.
- **OLAP servers**
  - Relational OLAP (ROLAP): extended relational DBMS that maps operations on multidimensional data to standard relational operations.
  - Multidimensional OLAP (MOLAP): special purpose server that directly implements multidimensional data and operations.
- **Clients**
  - Query and reporting tools
  - Analysis tools
  - Data mining tools (e.g., trend analysis, prediction)

Three Data Warehouse Models

- **Enterprise warehouse**
  - It contains detailed data as well as summarized data.
  - The enterprise warehouse collects all the information all the subjects spanning the entire organization
  - This provide us the enterprise-wide data integration.
  - The data is integrated from operational systems and external information providers.
  - This information can vary from a few gigabytes to hundreds of gigabytes, terabytes or beyond.

Data Mart

- a subset of corporate-wide data that is of value to a specific groups of users.
- in other words we can say that data mart contains only that data which is specific to a particular group.
- Its scope is confined to specific, selected groups, such as marketing data mart
- For example the marketing data mart may contain only data related to item, customers and sales.

Virtual warehouse

- A set of views over operational databases
- Only some of the possible summary views may be materialized.
- The view over a operational data warehouse is known as virtual warehouse.
- Building the virtual warehouse requires excess capacity on operational database servers.

Data Warehouse Usage

- Three kinds of data warehouse applications
  - Information processing
  - supports querying, basic statistical analysis, tables, charts and graphs
  - Analytical processing
  - multidimensional analysis of data warehouse data
  - supports basic OLAP operations, slice-dice, drilling, pivoting
  - Data mining
  - knowledge discovery from hidden patterns
  - supports associations, constructing analytical models, performing classification and prediction, and presenting the mining results using visualization tools

Other Names of Data Warehouse

DWH Examples

Airline:
In the Airline system, it is used for operation purpose like crew assignment, analyses of route profitability, frequent flyer program promotions, etc.

Banking:
It is widely used in the banking sector to manage the resources available on desk effectively. Few banks also used for the market research, performance analysis of the product and operations.

Healthcare:
Healthcare sector also used Data warehouse to strategize and predict outcomes, generate patient's treatment reports,
share data with tie-in insurance companies, medical aid services, etc.

Applications

Sr.No	Tasks	Deliverables
1	Need to define project scope	Scope Definition
2	Need to determine business needs	Logical Data Model
3	Define Operational Datastore requirements	Operational Data Store Model
4	Acquire or develop Extraction tools	Extract tools and Software
5	Define Data Warehouse Data requirements	Transition Data Model
6	Document missing data	To Do Project List
7	Maps Operational Data Store to Data Warehouse	D/W Data Integration Map
8	Develop Data Warehouse Database design	D/W Database Design
9	Extract Data from Operational Data Store	Integrated D/W Data Extracts
10	Load Data Warehouse	Initial Data Load
11	Maintain Data Warehouse	On-going Data Access and Subsequent Loads

DWH tools

Data Warehouse Schemas

- Data warehouses typically have schemas that are designed for data analysis, using tools such as OLAP tools.

- Thus, the data are usually multidimensional data, with **dimension attributes and measure attributes.**

- Tables containing multidimensional data are called **fact tables** and are usually very large.

- **A table recording sales** information for a retail store, with one tuple for each item that is sold, is a typical example of a fact table.

- **The dimensions of the *sales table*** *would include what the* item is (usually an item identifier such as that used in bar codes), the date when the item is sold, which location (store) the item was sold from, which customer bought the item, and so on. **The measure attributes** may include the number of items sold and the price of the items.

- To minimize storage requirements, dimension attributes are usually short identifiers that are foreign keys into other tables called **dimension tables.**

- **For** instance, a **fact table sales** would have attributes item id, store id, customer id, and date, and **measure attributes** number and price.

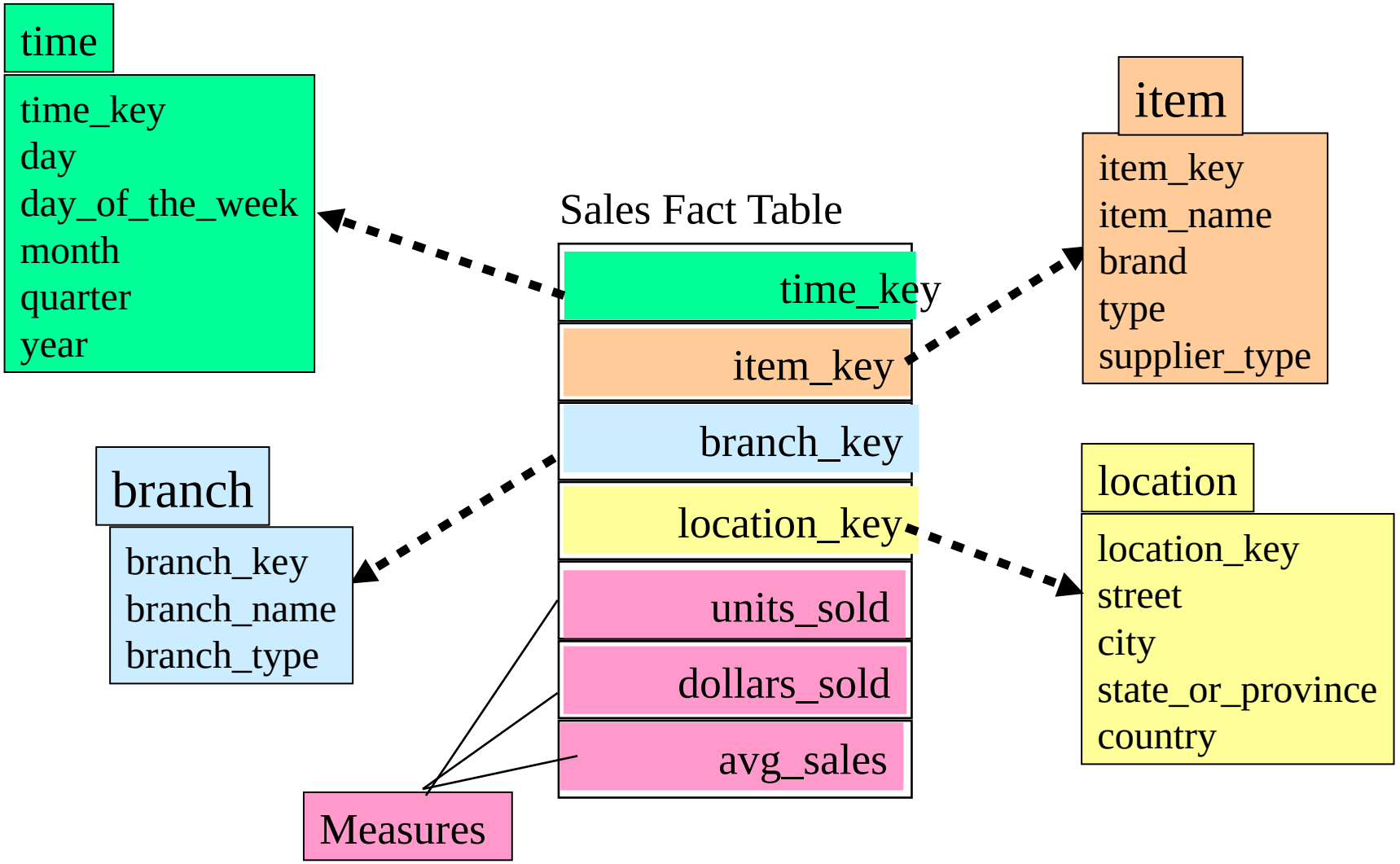
- The **attribute store id is a foreign key** into a dimension table store, which has other attributes such as store location (city, state, country).

- The **item id attribute** of the sales table would be a foreign key into a dimension table item info, which would contain information such as the name of the item, the category to which the item belongs, and other item details such as color and size.

- The **customer id attribute** would be a foreign key into a customer table containing attributes such as name and address of the customer.

- Similarly **the date attribute as a foreign key** into a date info table giving the month, quarter, and year of each date.

Example of Star Schema



•Such a schema, with a fact table, multiple dimension tables, and foreign keys from the fact table to the dimension tables, is called a **star schema.**

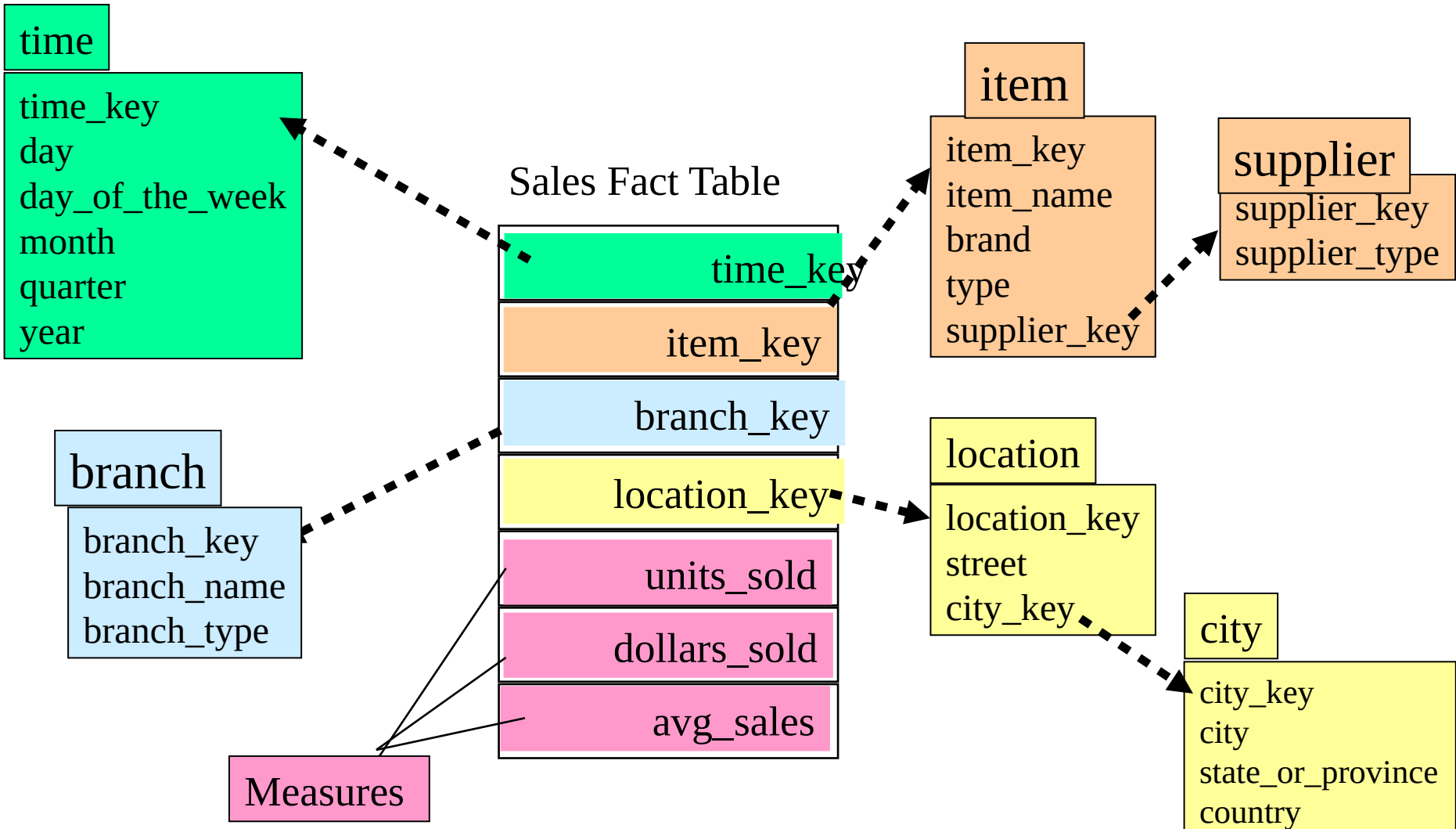
•More complex data-warehouse designs may have **multiple levels of dimension tables;**

•**For example**, the **item info table** may have an attribute manufacturer id that is a foreign key into another table giving details of the manufacturer.

•Such schemas are called **snowflake schemas.**

•**Complex data warehouse** designs may also have more than one fact table.

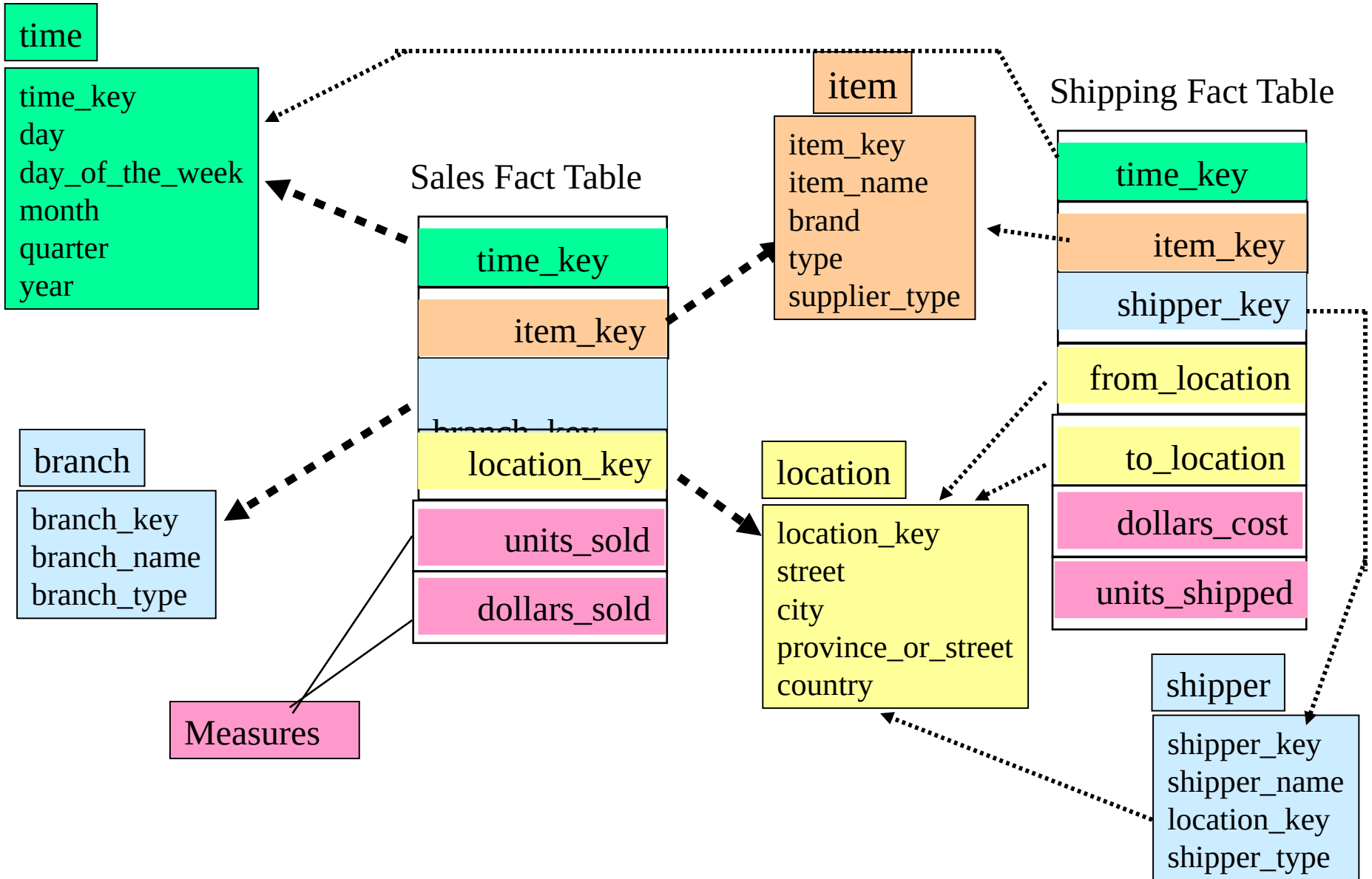
Example of Snowflake Schema



Fact Constellation

- Multiple fact tables share dimension tables, viewed as a collection of stars, therefore called galaxy schema or fact constellation

Example of Fact Constellation



Can u design a schema for Academic schedule and the classroom/Laboratory utilization of a college

Average class room utilization
Average class room utilization per semester
Maximum utilization per semester
Minimum utilization of a class room per semester
Minimum utilization of a class room per branch
Minimum utilization of a class room per class(FY to Mtech)
Average availability in hours of a seminar hall
Average availability in morning slot(1st quartile) of a seminar hall
Average availability during lunch break of a seminar hall etc

2 dimensional relation table :
1. class room and time, (derived attribute : total utilization of classrooms, Weekly utilization,monthly yearly utilization)
2. class room and year (total class rooms allotted to FY/SY/TY and so on, max Class rooms allotted to which year etc)
3. class room and branch

Managerial decision

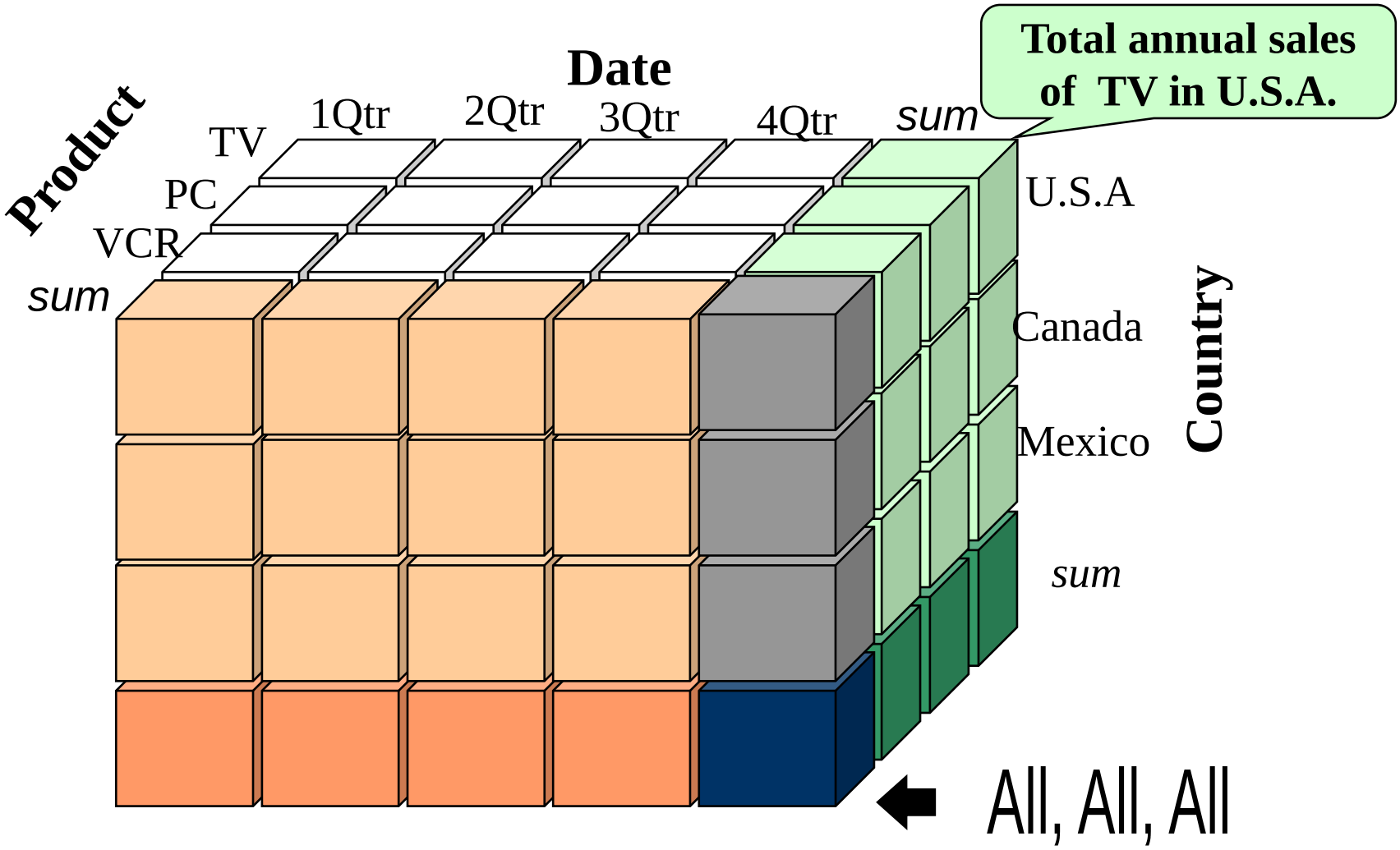
Conceptual Modeling of Data Warehouses

- Modeling data warehouses: dimensions & measures
  - Star schema: A fact table in the middle connected to a set of dimension tables
  - Snowflake schema: A refinement of star schema where some dimensional hierarchy is normalized into a set of smaller dimension tables.
  - Fact constellations: Multiple fact tables share dimension tables, viewed as a collection of stars, therefore called galaxy schema or fact constellation

Multidimensional Data Model

- Database is a set of *facts* (points) in a multidimensional space.
- A fact has a *measure* dimension
  - quantity that is analyzed, e.g., sale, budget
- A set of *dimensions* on which data is analyzed
  - e.g. , store, product, date associated with a sale amount
- Each dimension has a set of *attributes*
  - e.g., owner city and region of store
- Attributes of a dimension may be related by partial order
  - *Hierarchy*: e.g., street > region >city
  - *Lattice/pattern*: e.g., date> month>year, date>week>year

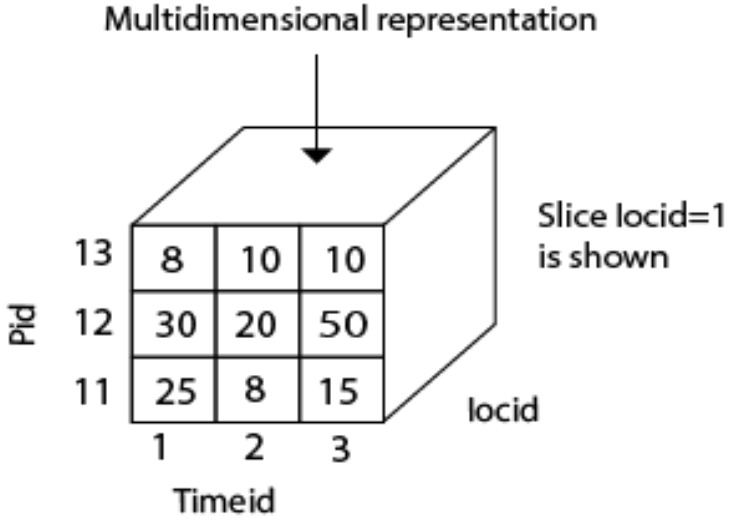
Multidimensional Data



Example

Tabular representation

Pid	Timeid	locid	Sales
11	1	1	25
11	2	1	8
11	3	1	15
12	1	1	30
12	2	1	20
12	3	1	50
13	1	1	8
13	2	1	10
13	3	1	10
11	1	2	35



OLAP

- OLAP technology is a vast improvement over traditional relational database management systems (RDBMS).

- Relational databases, which have a two-dimensional structure, do not allow the multidimensional data views that OLAP provides.

- Traditionally used as an analytical tool for marketing and financial reporting, OLAP is viewed as a valuable tool for any management system that needs to create a flexible decision support system.

- Traditional tools of report writers, query products, spreadsheets, & language interfaces do not match the user expectations as far as performing multidimensional analysis with complex calculations is concerned.
- Tools used with OLTP and basic DW environments do not match up to the task.
- OLAP provides the multidimensional capabilities that most organizations need today.

OLAP

- OLAP is a category of software technology that enables analysts, managers, and executives to gain insight into the data through fast, consistent, interactive, access in a wide variety of possible views of information that has been transformed from raw data to reflect the real dimensionality of the enterprise as understood by the user.

What and Why OLAP?

- OLAP is the dynamic synthesis, analysis, and consolidation of large volumes of multi-dimensional data.
- OLAP is the term that describes a technology that uses multi-dimensional view of aggregate data to provide quick access to strategic information for the purposes of advanced analysis.
- OLAP enables users to gain a deeper understanding and knowledge about various aspects of their corporate data through fast, consistent, interactive access to a variety of possible views of data.

OLAP Applications

- **Finance:** Budgeting, activity-based costing, financial performance analysis, and financial modeling.
- **Sales:** Sales analysis and sales forecasting.
- **Marketing:** Market research analysis, sales forecasting, promotions analysis, customer analysis, and market/customer segmentation.
- **Manufacturing:** Production planning and defect analysis.

OLAP Key Features

- Multi-dimensional views of data.
- Support for complex calculations.
- Time Intelligence.

Representation of Multi-Dimensional Data

- OLAP database servers use multi-dimensional structures to store data and relationships between data.
- Multi-dimensional structures are best-visualized as cubes of data, and cubes within cubes of data. Each side of a cube is a dimension.

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

(a)

City

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

(b)

Representation of Multi-Dimensional Data

- Multi-dimensional databases are a compact and easy-to-understand way of visualizing and manipulating data elements that have many inter-relationships.

- The cube can be expanded to include another dimension, for example, the number of sales staff in each city.

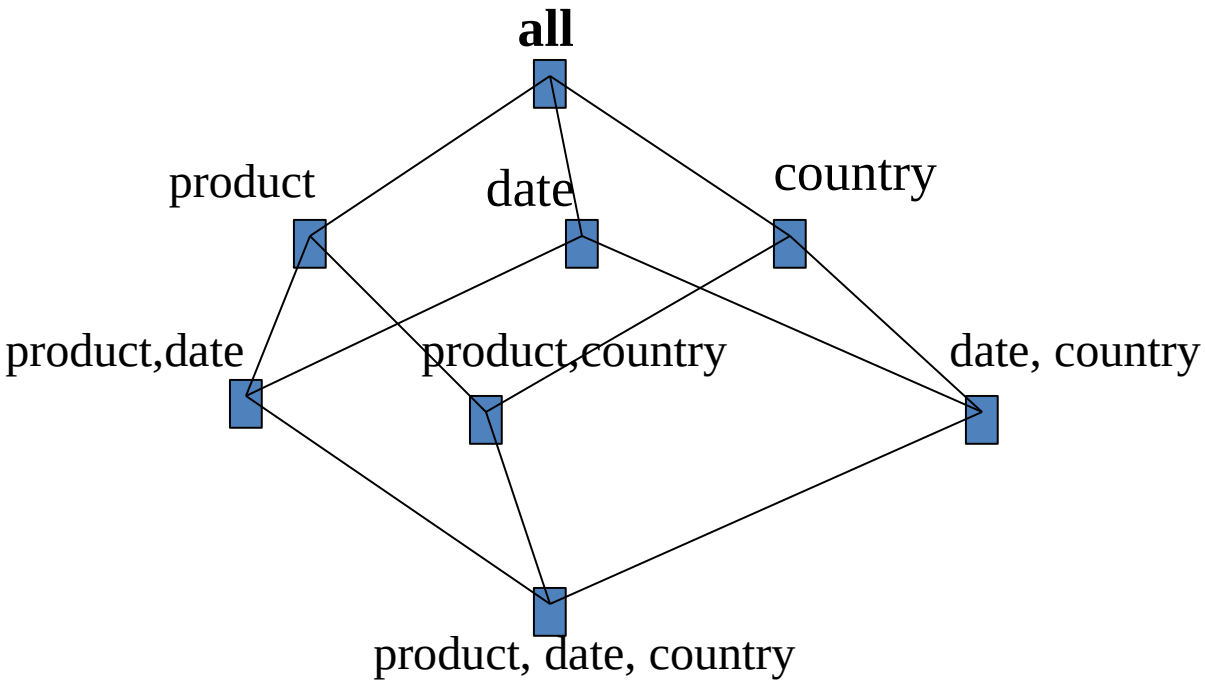
- The response time of a multi-dimensional query depends on how many cells have to be added on-the-fly.

- As the number of dimensions increases, the number of cube's cells increases exponentially.

From Tables and Spreadsheets to Data Cubes

- A data warehouse 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
  - Dimension tables, such as item (item_name, brand, type), or time(day, week, month, quarter, year)
  - Fact table contains measures (such as dollars_sold) and keys to each of the related dimension tables
- 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.

Cuboids Corresponding to the Cube



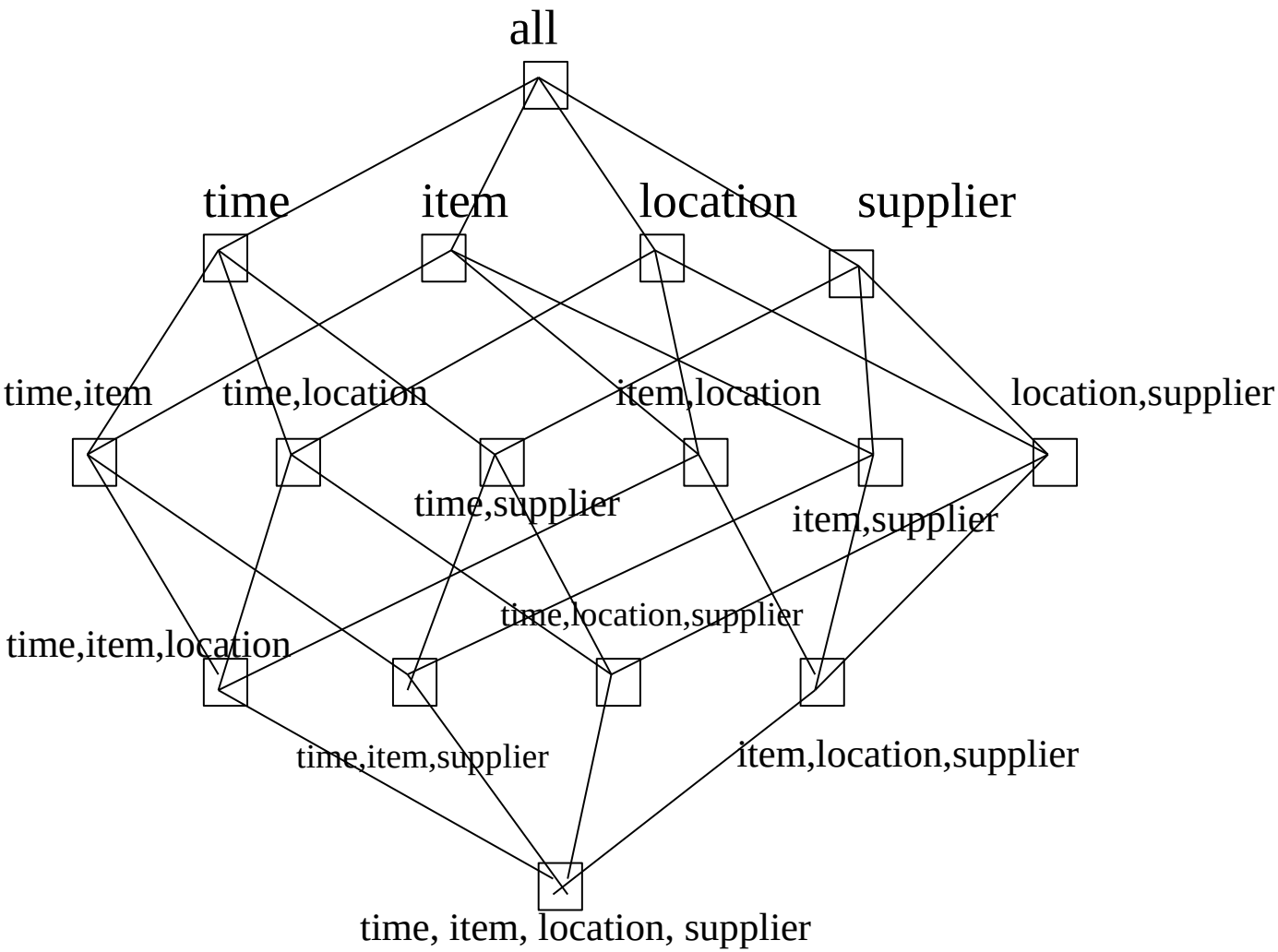
0-D(apex) cuboid

1-D cuboids

2-D cuboids

3-D(base) cuboid

Lattice of Cuboids



0-D(apex) cuboid

1-D cuboids

2-D cuboids

3-D cuboids

4-D(base) cuboid

OLAP Tools - Categories

- OLAP tools are categorized according to the architecture used to store and process multi-dimensional data.

- There are four main categories of OLAP tools as defined by Berson and Smith (1997) and Pends and Greeth (2001) including:
  - Multi-dimensional OLAP (MOLAP)
  - Relational OLAP (ROLAP)
  - Hybrid OLAP (HOLAP)
  - Specialized SQL Servers

Relational OLAP(ROLAP)

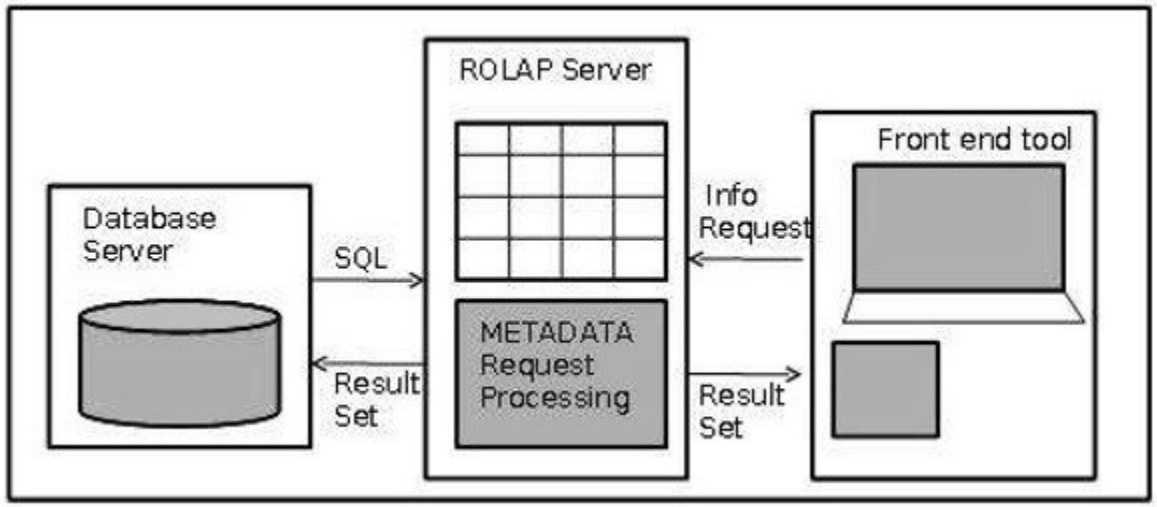
- The Relational OLAP servers are placed between relational back-end server and client front-end tools.
- To store and manage warehouse data the Relational OLAP use relational or extended-relational DBMS.
- ROLAP is the fastest-growing type of OLAP tools.

Relational OLAP(ROLAP)

- The Relational OLAP servers are placed between relational back-end server and client front-end tools.
- To store and manage warehouse data the Relational OLAP use relational or extended-relational DBMS.
- ROLAP is the fastest-growing type of OLAP tools.

Relational OLAP(ROLAP)

ROLAP servers can be easily used with existing RDBMS.
Data can be stored efficiently, since no zero facts can be stored.
ROLAP tools do not use pre-calculated data cubes.



Multidimensional OLAP (MOLAP)

•Multidimensional OLAP (MOLAP) uses the array-based multidimensional storage engines for multidimensional views of data.

•With multidimensional data stores, the storage utilization may be low if the data set is sparse.

•Therefore many MOLAP Server uses the two level of data storage representation to handle dense and sparse data sets.

•**Advantages**

•MOLAP allows fastest indexing to the pre-computed summarized data.

•Helps the users connected to a network who need to analyze larger, less-defined data.

•Easier to use, therefore MOLAP is suitable for inexperienced users.

Disadvantages

MOLAP are not capable of containing detailed data.

The storage utilization may be low if the data set is sparse.

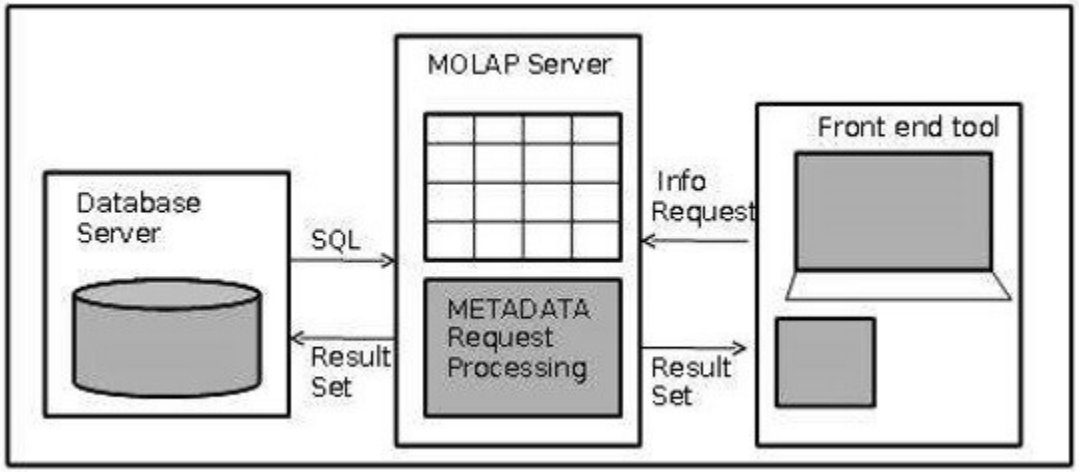
•MOLAP tools use specialized data structures and multi-dimensional database management systems (MDDBMS) to organize, navigate, and analyze data.

•MOLAP data structures use array technology and efficient storage techniques that minimize the disk space requirements through sparse data management.

•The development issues associated with MOLAP:

- Only a limited amount of data can be efficiently stored and analyzed.
- Navigation and analysis of data are limited because the data is designed according to previously determined requirements.
- MOLAP products require a different set of skills and tools to build and maintain the database.

- MOLAP tools process information with consistent response time regardless of level of summarizing or calculations selected.
- MOLAP tools need to avoid many of the complexities of creating a relational database to store data for analysis.
- MOLAP tools need fastest possible performance.
- MOLAP server adopts two level of storage representation to handle dense and sparse data sets.
- Denser sub-cubes are identified and stored as array structure.
- Sparse sub-cubes employ compression technology.



Hybrid OLAP (HOLAP)

- The hybrid OLAP technique combination of ROLAP and MOLAP both.
- It has both the higher scalability of ROLAP and faster computation of MOLAP.
- HOLAP server allows to store the large data volumes of detail data.
- HOLAP tools deliver selected data directly from DBMS or via MOLAP server to the desktop (or local server) in the form of data cube, where it is stored, analyzed, and maintained locally is the fastest-growing type of OLAP tools.

With this use of the two OLAPs, the data is stored in both multidimensional databases and relational databases. The decision to access one of the databases depends on which is most appropriate for the requested processing application or type.

https://www.sisense.com/glossary/olap/

- Like MOLAP, HOLAP causes the aggregations of the partition to be stored in a multidimensional structure in an SQL Server Analysis Services instance.

- HOLAP does not cause a copy of the source data to be stored.

- For queries that access only summary data in the aggregations of a partition, HOLAP is the equivalent of MOLAP.

-

➢Specialized SQL Servers

- specialized SQL servers provides advanced query language and query processing support for SQL queries over star and snowflake schemas in a read-only environment.

OLAP Operations

- Roll-Up
- Drill-Down
- Slice & Dice
- Pivot
- Drill-Across
- Drill-Through

Roll-Up Operation

❑This operation performs aggregation on a data cube in any of the following way:

•By climbing up a concept hierarchy for a dimension

•By dimension reduction.

✓The roll-up operation is performed by climbing up a concept hierarchy for the dimension location.

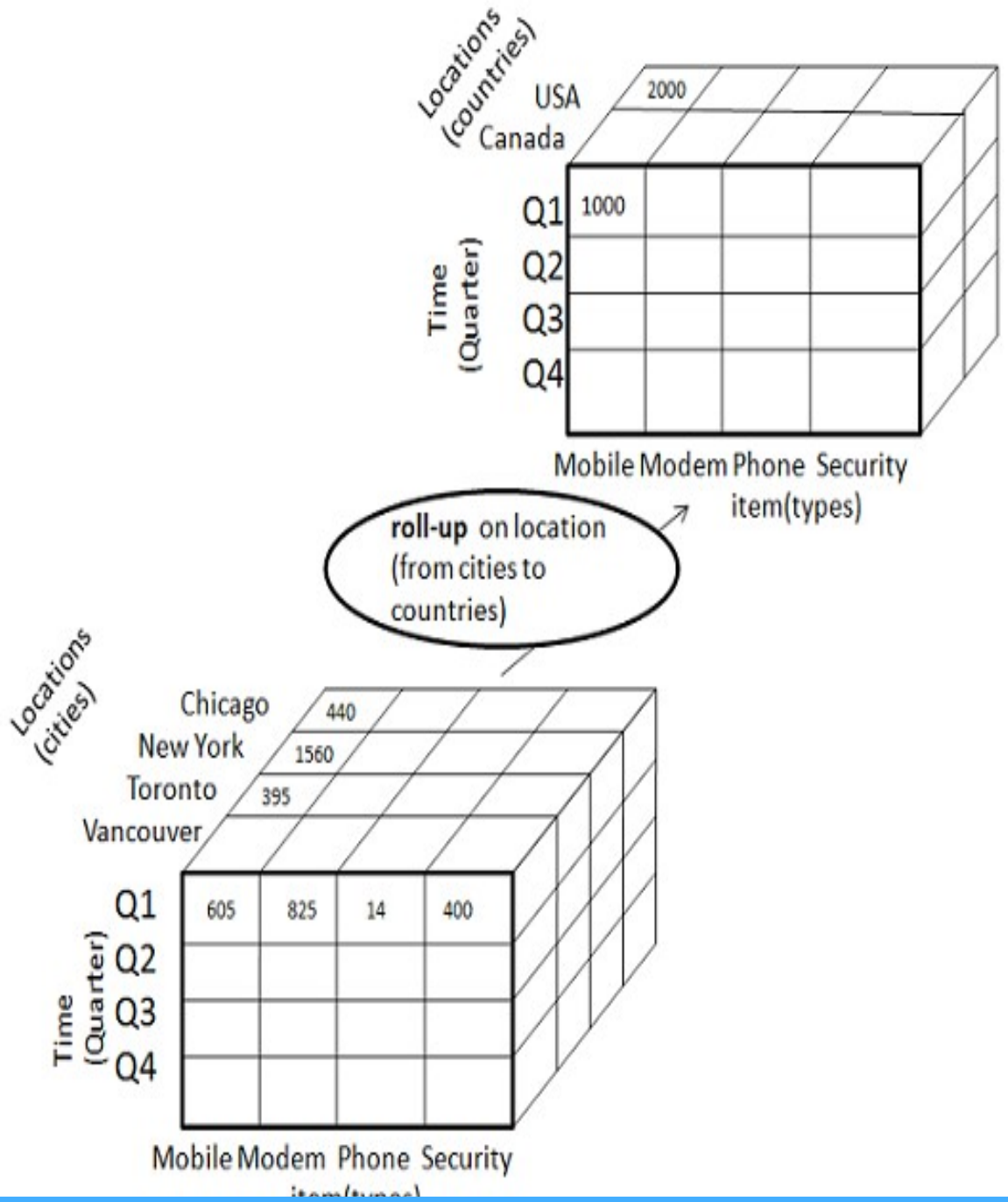
✓Initially the concept hierarchy was “street < city < state < country”.

✓On rolling up the data is aggregated by ascending the location hierarchy from the level of city to level of country.

✓The data is grouped into cities rather than countries.

✓When roll-up operation is performed then one or more dimensions from the data cube are removed.

•Consider the diagram showing the roll-up operation


Locations (countries)
USA
Canada
2000
1000
Q1
Q2
Q3
Q4
Time (Quarter)
Mobile Modem Phone Security
item(types)
roll-up on location (from cities to countries)
Locations (cities)
Chicago
New York
Toronto
Vancouver
440
1560
395
605
825
14
400
Q1
Q2
Q3
Q4
Time (Quarter)
Mobile Modem Phone Security

- When roll-up is performed by dimension reduction, one or more dimensions are removed from the given cube.

- For example, consider the sale data cube containing only two dimensions like location and time.

- Roll-up may be performed by removing the time dimension, resulting in an aggregation of the total sales by location, rather than by location and by time.

Drill-down

❑Drill-down operation is reverse of the roll-up. This operation is performed by either of the following way:

•By stepping down a concept hierarchy for a dimension.

•By introducing new dimension.

✓The drill-down operation is performed by stepping down a concept hierarchy for the dimension time.

✓Initially the concept hierarchy was "day < month < quarter < year.”

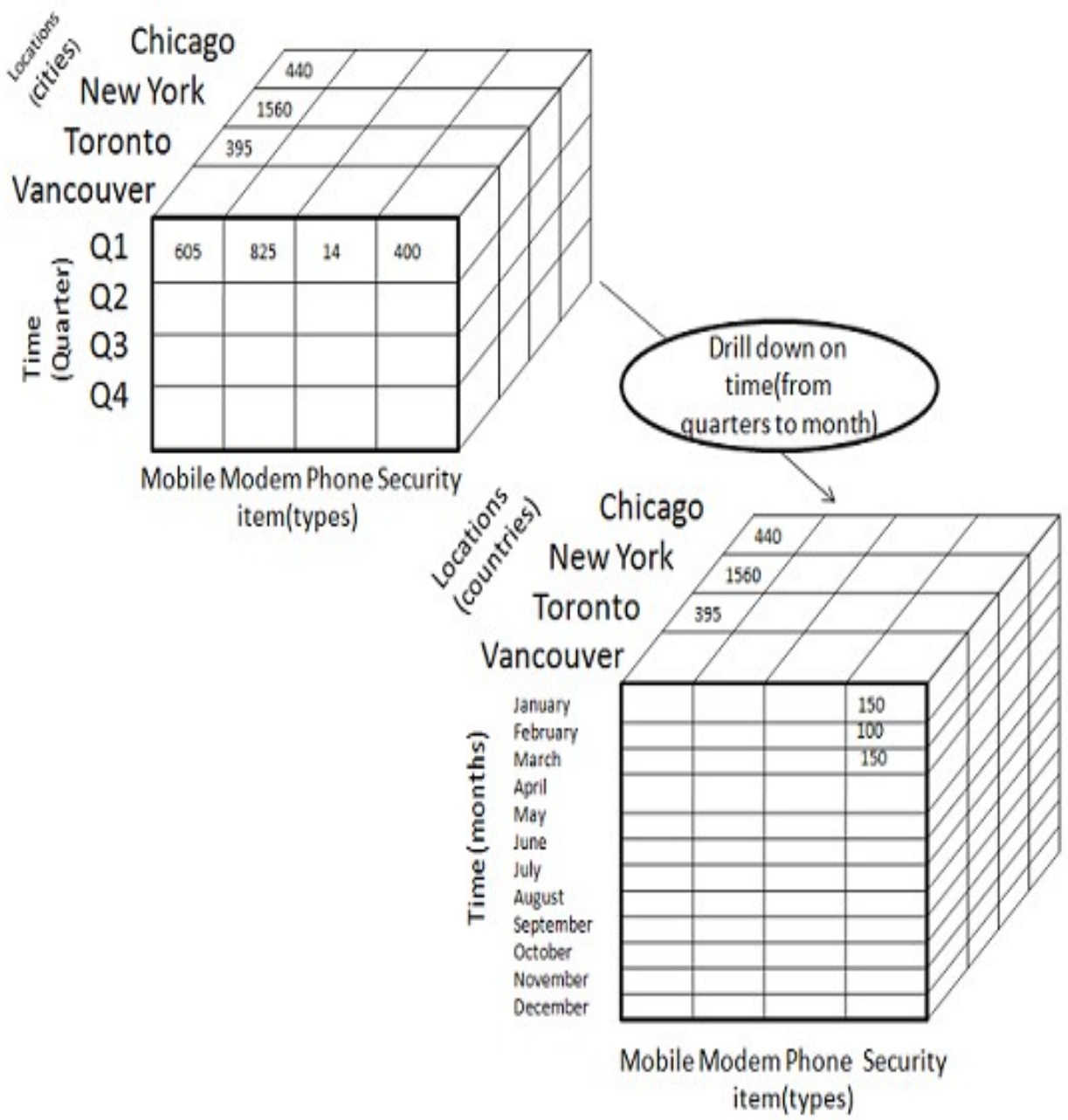
✓On drill-up the time dimension is descended from the level quarter to the level of month.

✓When drill-down operation is performed then one or more dimensions from the data cube are added.

✓It navigates the data from less detailed data to highly detailed data.

❑Consider the diagram showing the drill-down operation:

Drill-down

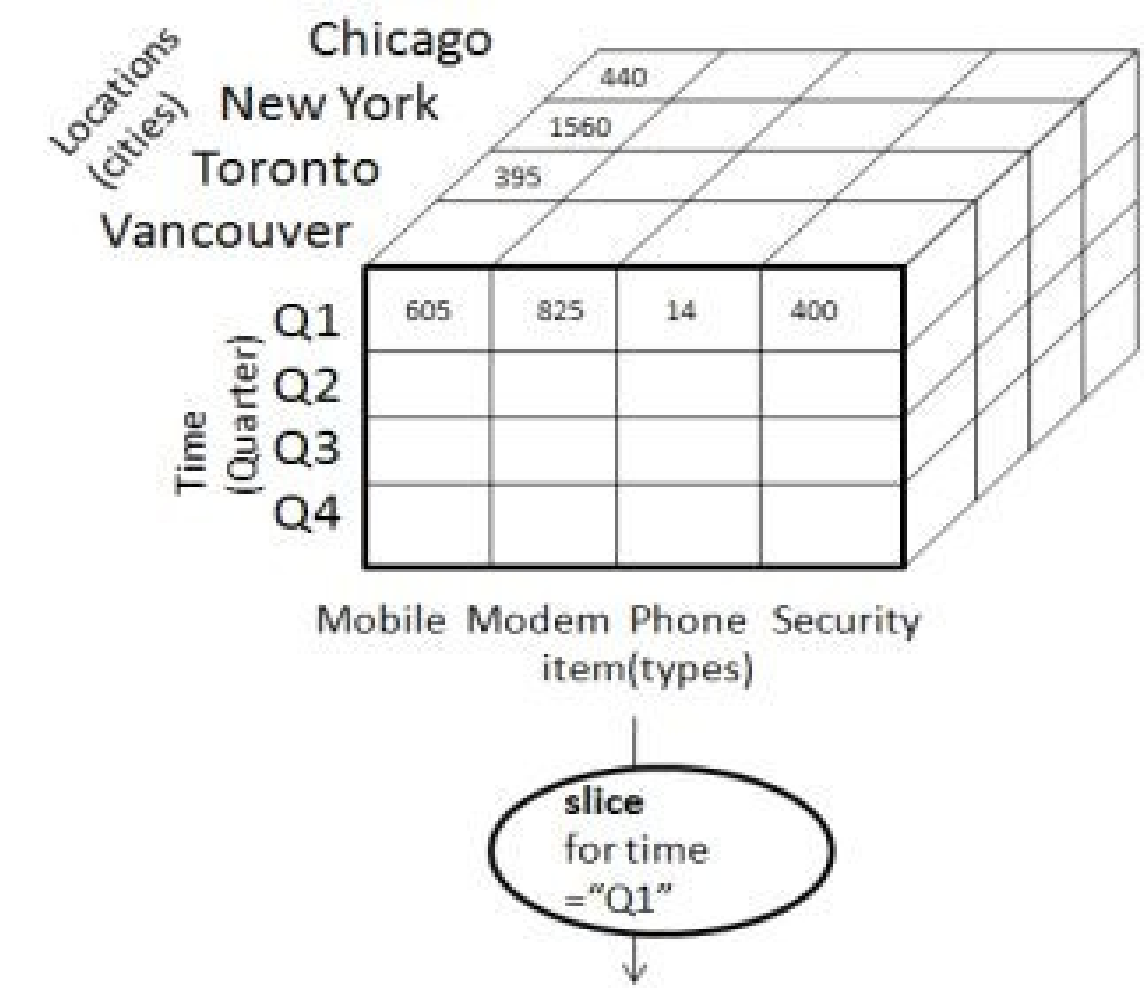

Locations (cities)
Chicago
New York
Toronto
Vancouver
440
1560
395
Time (Quarter)
Q1
Q2
Q3
Q4
605
825
14
400
Mobile Modem Phone Security
item(types)
Drill down on time(from quarters to month)
Locations (countries)
Chicago
New York
Toronto
Vancouver
440
1560
395
Time (months)
January
February
March
April
May
June
July
August
September
October
November
December
150
100
150
Mobile Modem Phone Security
item(types)

- The resulting data cube details the total sales per month rather than summarizing them by quarter.

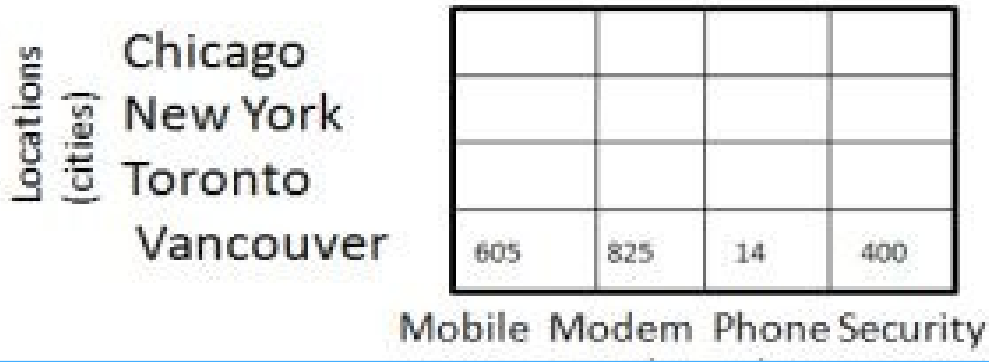
- Because a drill down adds more detail to the given data, it can also be performed by adding new dimensions to a cube.

Slice Operation

- The slice operation performs selection of one dimension on a given cube and give us a new sub cube.
- The Slice operation is performed for the dimension time using the criterion time ="Q1".
- It will form a new sub cube by selecting one or more dimensions.
- Consider the diagram showing the slice operation.


Locations (cities)
Chicago
New York
Toronto
Vancouver
440
1560
395
Time (Quarter)
Q1
Q2
Q3
Q4
605
825
14
400
Mobile
Modem
Phone
Security
item(types)

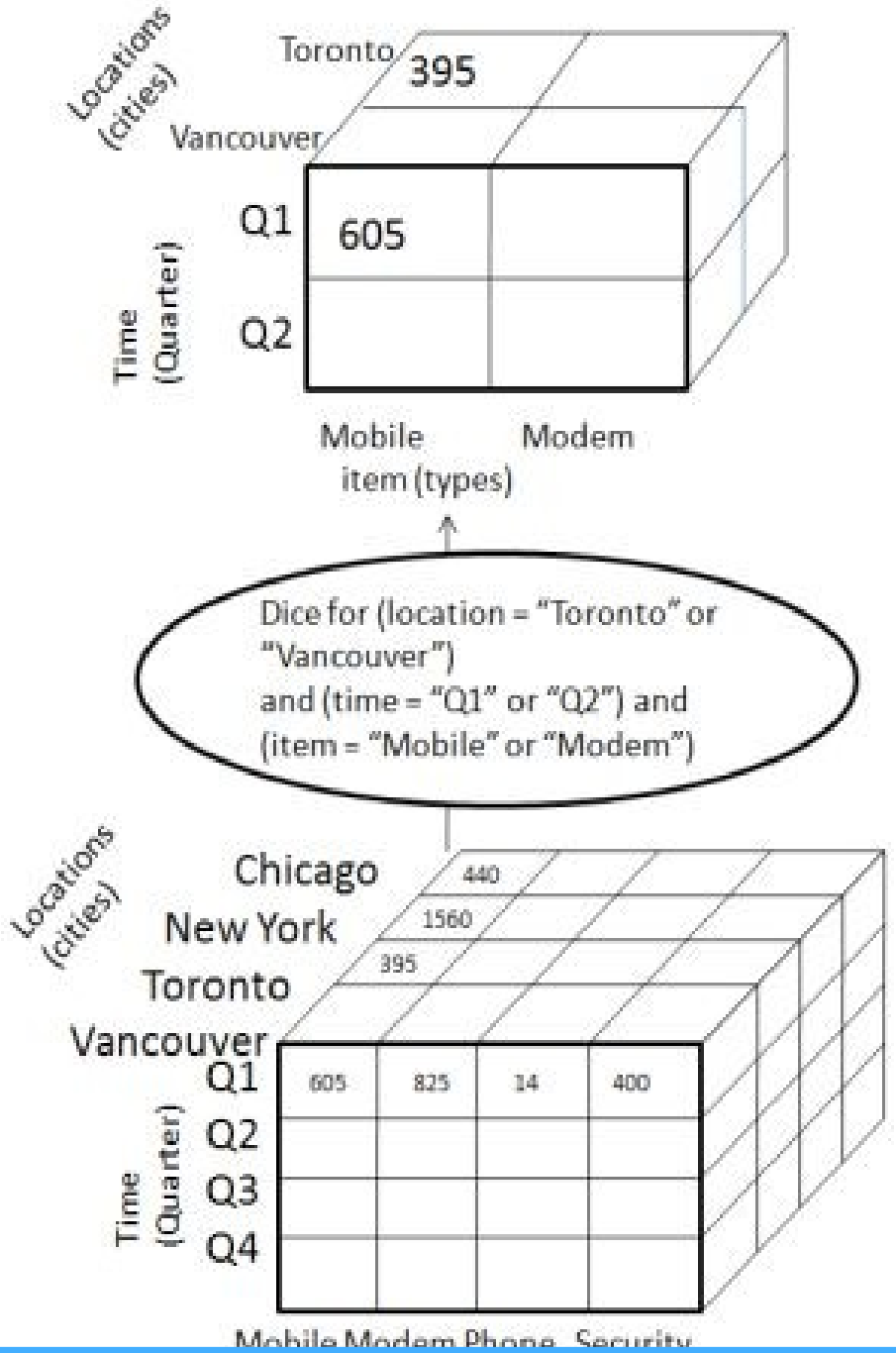
slice
for time
="Q1"


Locations (cities)
Chicago
New York
Toronto
Vancouver
605
825
14
400
Mobile
Modem
Phone
Security

Dice Operation

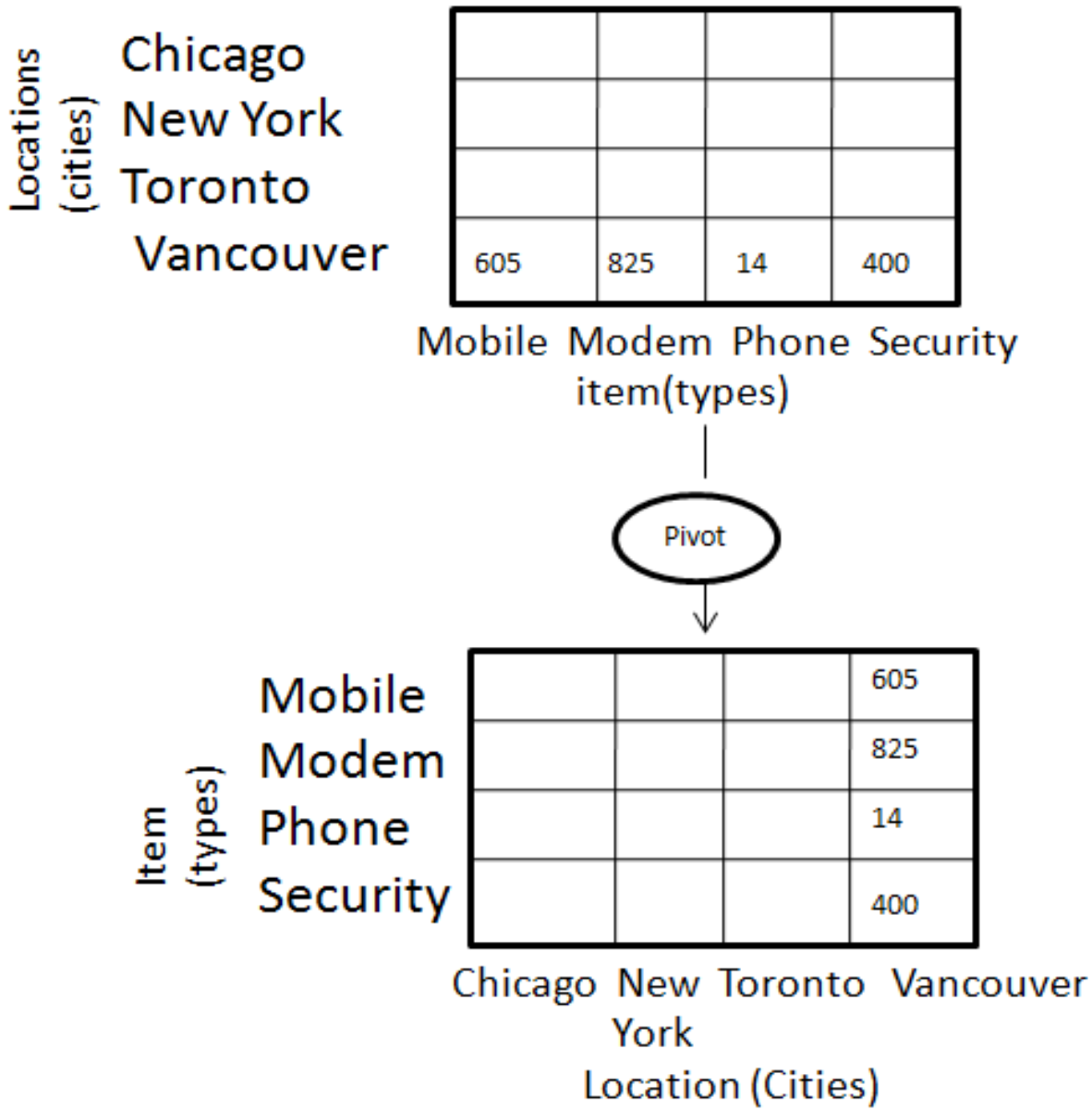
- The Dice operation performs selection of two or more dimension on a given cube and give us a new subcube.
- The dice operation on the cube based on the following selection criteria that involve three dimensions.

✓ (location = "Toronto" or "Vancouver“)
✓ (time = "Q1" or "Q2“)
✓ (item =" Mobile" or "Modem").


Locations (cities)
Toronto
395
Vancouver
Q1
605
Q2
Time (Quarter)
Mobile
Modem
item (types)
Dice for (location = "Toronto" or "Vancouver")
and (time = "Q1" or "Q2") and
(item = "Mobile" or "Modem")
Locations (cities)
Chicago
440
New York
1560
Toronto
395
Vancouver
Q1
605
825
14
400
Q2
Q3
Q4
Time (Quarter)
Mobile Modem Phone Security

Pivot Operation

- The pivot operation is also known as rotation.
- It rotates the data axes in view in order to provide an alternative presentation of data.
- Consider the diagram showing the pivot operation.


Locations (cities)
Chicago
New York
Toronto
Vancouver
605
825
14
400
Mobile Modem Phone Security
item(types)
Pivot
Item (types)
Mobile
Modem
Phone
Security
605
825
14
400
Chicago New York Toronto Vancouver
Location (Cities)

Other OLAP Operations

- **Drill-Across:** Queries involving more than one fact table.

- **Drill-Through:** Makes use of SQL to drill through the bottom level of a data cube down to its back-end relational tables.

- **Pivot (rotate):** Pivot (also called **"rotate")** is a visualization operation which rotates the data axes in view in order to provide an alternative presentation of the data.

Data Warehouse vs. Operational DBMS

- **OLTP (on-line transaction processing)**
  - Major task of traditional relational DBMS
  - Day-to-day operations: purchasing, inventory, banking, manufacturing, registration, accounting, etc.
- **OLAP (on-line analytical processing)**
  - Major task of data warehouse system
  - Data analysis and decision making

➢**User and system orientation:** customer vs. market:

- An OLTP is custom-oriented and is used for transaction and query processing by clerks, clients and information technology.
- An OLAP is market oriented and is used for data analysis by knowledge workers, including managers,executive and analysis.

➢**Data contents:**

- An OLTP manages current data that are too detailed and can be easily used for decision making.
- An OLAP system manages large amount of historical data,provide facilities for summarization.

➢**Database design**

- An OLTP uses ER-model and application oriented database design and an OLAP uses subject oriented database system.

➢**View**

- An OLTP focuses on current data while OLAP focuses on historical data.

Parameters	OLTP	OLAP
Characteristics	Operational processing	Informational processing
Orientation	Transaction	Analysis
users	clerk, IT & database professional	knowledge worker(manager)
function	day to day operations	decision support
DB design	application-oriented, ER-based	subject-oriented(star, snowflake)
data	current, up-to-date detailed,	historical, summarized, multidimensional
access	read/write index/hash on primary key	lots of scans, mostly read
unit of work	short, simple transaction	complex query
No.of users	thousands	Hundreds
DB size	100 MB to GB	100 GB to TB

Parameters	OLTP	OLAP
Characteristic	It is characterized by large numbers of short online transactions.	It is characterized by a large volume of data.
Functionality	OLTP is an online database modifying system.	OLAP is an online database query management system.
Method	OLTP uses traditional DBMS.	OLAP uses the data warehouse.
Query	Insert, Update, and Delete information from the database.	Mostly select operations
Table	Tables in OLTP database are normalized.	Tables in OLAP database are **not** normalized.
Source	OLTP and its transactions are the sources of data.	Different OLTP databases become the source of data for OLAP.
Data Integrity	OLTP database must maintain data integrity constraint.	OLAP database does not get frequently modified. Hence, data integrity is not an issue.
Response time	It's response time is in millisecond.	Response time in seconds to minutes.

Parameters	OLTP	OLAP
Usefulness	It helps to control and run fundamental business tasks.	It helps with planning, problem-solving, and decision support.
Operation	Allow read/write operations.	Only read and rarely write.
Query Type	Queries in this process are standardized and simple.	Complex queries involving aggregations.
Design	DB design is application oriented. Example: Database design changes with industry like Retail, Airline, Banking, etc.	DB design is subject oriented. Example: Database design changes with subjects like sales, marketing, purchasing, etc.

Parameters	OLTP	OLAP
User type	It is used by Data critical users like clerk, DBA & Data Base professionals.	Used by Data knowledge users like workers, managers, and CEO.
Purpose	Designed for real time business operations.	Designed for analysis of business measures by category and attributes.
Performance metric	Transaction throughput is the performance metric	Query throughput is the performance metric.
Number of users	This kind of Database users allows thousands of users.	This kind of Database allows only hundreds of users.
Productivity	It helps to Increase user's self-service and productivity	Help to Increase productivity of the business analysts.
Data quality	The data in the OLTP database is always detailed and organized.	The data in OLAP process might not be organized.

Parameters	OLTP	OLAP
Challenge	Data Warehouses historically have been a development project which may prove costly to build.	An OLAP cube is not an open SQL server data warehouse. Therefore, technical knowledge and experience is essential to manage the OLAP server.
Process	It provides fast result for daily used data.	It ensures that response to the query is quicker consistently.
Characteristic	It is easy to create and maintain.	It lets the user create a view with the help of a spreadsheet.
Style	OLTP is designed to have fast response time, low data redundancy and is normalized.	A data warehouse is created uniquely so that it can integrate different data sources for building a consolidated database

Big Data Lakes

- A **data lake** is a storage repository that holds a vast amount of raw **data** in its native format until it is needed.

- While a hierarchical **data** warehouse stores **data** in files or folders, a **data lake** uses a flat architecture to store **data**.

Big Data Lakes

- During the development of a data warehouse, a considerable amount of time is spent analyzing data sources, understanding business processes and profiling data. The result is a highly structured data model designed for reporting.

- A large part of this process includes making decisions about what data to include and to not include in the warehouse. Generally, if data isn't used to answer specific questions or in a defined report, it may be excluded from the warehouse.

Big Data Lakes

- This is usually done to simplify the data model and also to conserve space on expensive disk storage that is used to make the data warehouse performant.

- In contrast, the data lake retains ALL data. Not just data that is in use today but data that may be used and even data that may never be used just because it MIGHT be used someday. Data is also kept for all time so that we can go back in time to any point to do analysis.

Big Data Lakes

➢ **What is Data Lake?**

- A Data Lake is a storage repository that can store large amount of structured, semi-structured, and unstructured data. It is a place to store every type of data in its native format with no fixed limits on account size or file.

- It offers high data quantity to increase analytic performance and native integration.

- Data Lake is like a large container which is very similar to real lake and rivers. Just like in a lake you have multiple tributaries coming in, a data lake has structured data, unstructured data, machine to machine, logs flowing through in real-time.

Big Data Lakes

Big Data Lakes

- **Key points of Data Lake**

Data Ingestion

Data Ingestion allows connectors to get data from a different data sources and load into the Data lake.

Data Ingestion supports: All types of Structured, Semi-Structured, and Unstructured data.

Multiple ingestions like Batch, Real-Time, One-time load.

Many types of data sources like Databases, Webservers, Emails, IoT etc.

Data Storage

Data storage should be scalable, offers cost-effective storage and allow fast access to data exploration. It should support various data formats.

Data Governance

Data governance is a process of managing availability, usability, security, and integrity of data used in an organization.

Big Data Lakes

Security

Security needs to be implemented in every layer of the Data lake. It starts with Storage, Detection, and Consumption. The basic need is to stop access for unauthorized users. It should support different tools to access data with easy to navigate GUI and Dashboards.

Authentication, Accounting, Authorization and Data Protection are some important features of data lake security.

Data Quality:

Data quality is an essential component of Data Lake architecture. Data is used to extract business value. Extracting insights from poor quality data will lead to poor quality insights.

Data Discovery

Data Discovery is another important stage before you can begin preparing data or analysis. In this stage, tagging technique is used to express the data understanding, by organizing and interpreting the data ingested in the Data lake.

Big Data Lakes

Data Auditing

Two major Data auditing tasks are tracking changes to the key dataset.
Tracking changes to important dataset elements
Captures how/ when/ and who changes to these elements.
Data auditing helps to evaluate risk and compliance.

Data Lineage

This component deals with data's origins. It mainly deals with where it movers over time and what happens to it. It eases errors corrections in a data analytics process from origin to destination.

Data Exploration

It is the beginning stage of data analysis. It helps to identify right dataset is vital before starting Data Exploration.
All given components need to work together to play an important part in Data lake building easily evolve and explore the environment.

Big Data Lakes

➢ **Reasons for using Data Lake are:**

- With the increase in data volume, data quality, and metadata, the quality of analyses also increases.
- Data Lake offers business Agility
- Machine Learning and Artificial Intelligence can be used to make profitable predictions.
- It offers a competitive advantage to the implementing organization.

Big Data Lakes

Characteristics	Data Warehouse	Data Lake
Data	Relational from transactional systems, operational databases, and line of business applications	Non-relational and relational from IoT devices, web sites, mobile apps, social media, and corporate applications
Schema	Designed prior to the DW implementation (schema-on-write)	Written at the time of analysis (schema-on-read)
Price/Performance	Fastest query results using higher cost storage	Query results getting faster using low-cost storage

Big Data Lakes

Characteristics	Data Warehouse	Data Lake
Data Quality	Highly curated data that serves as the central version of the truth	Any data that may or may not be curated (ie. raw data)
Users	Business analysts	Data scientists, Data developers, and Business analysts (using curated data)
Analytics	Batch reporting, BI and visualizations	Machine Learning, Predictive analytics, data discovery and profiling

Big Data Lakes

Parameters	Data Lakes	Data Warehouse
Data	Data lakes store everything.	Data Warehouse focuses only on Business Processes.
Processing	Data are mainly unprocessed	Highly processed data.
Type of Data	It can be Unstructured, semi-structured and structured.	It is mostly in tabular form & structure.
Task	Share data stewardship	Optimized for data retrieval
Agility	Highly agile, configure and reconfigure as needed.	Compare to Data lake it is less agile and has fixed configuration.
Users	Data Lake is mostly used by Data Scientist	Business professionals widely use data Warehouse
Storage	Data lakes design for low-cost storage.	Expensive storage that give fast response times are used

Big Data Lakes

Parameters	Data Lakes	Data Warehouse
Security	Offers lesser control.	Allows better control of the data.
Schema	Schema on reading (no predefined schemas)	Schema on write (predefined schemas)
Data Processing	Helps for fast ingestion of new data.	Time-consuming to introduce new content.
Data Granularity	Data at a low level of detail or granularity.	Data at the summary or aggregated level of detail.
Tools	Can use open source/tools like Hadoop/ Map Reduce	Mostly commercial tools.

Big Data Lakes

Summary:

- A Data Lake is a storage repository that can store large amount of structured, semi-structured, and unstructured data.
- The main objective of building a data lake is to offer an unrefined view of data to data scientists.
- Data Ingestion, Data storage, Data quality, Data Auditing, Data exploration, Data discover are some important components of Data Lake Architecture
- Design of Data Lake should be driven by what is available instead of what is required.
- Data Lake reduces long-term cost of ownership and allows economic storage of files
- The biggest risk of data lakes is security and access control. Sometimes data can be placed into a lake without any oversight, as some of the data may have privacy and regulatory need.

Case Study: Data Lakes

Amazon Simple Storage Service (S3) is the largest and most performant object storage service for structured and unstructured data and the storage service of choice to build a data lake. With Amazon S3, you can cost-effectively build and scale a data lake of any size in a secure environment where data is protected by 99.999999999% of durability.

With a data lake built on Amazon S3, you can use native AWS services to run big data analytics, artificial intelligence (AI), machine learning (ML), high-performance computing (HPC) and media data processing applications to gain insights from your unstructured data sets.

You also have the flexibility to use your preferred analytics, AI, ML, and HPC applications from the Amazon Partner Network (APN). Because Amazon S3 supports a wide range of features, IT managers, storage administrators, and data scientists are empowered to enforce access policies, manage objects at scale and audit activities across their S3 data lakes.

Amazon S3 hosts more than 10,000 data lakes.

References

- https://aws.amazon.com/big-data/datalakes-and-analytics/what-is-a-data-lake/
- https://www.guru99.com/data-lake-architecture.html
- https://aws.amazon.com/products/storage/data-lake-storage/

Thank You