I.T.S
GHAZIABAD

Data Warehousing & Mining

UNIT – I

I.T.S GHAZIABAD

Syllabus of Unit - I

- DSS-Uses, definition, Operational Database.
- Introduction to DATA Warehousing. Data-Mart,
- Concept of Data-Warehousing,
- Multi Dimensional Database Structures.
- Client/Server Computing Model & Data Warehousing
- Parallel Processors & Cluster Systems. Distributed DBMS implementations.

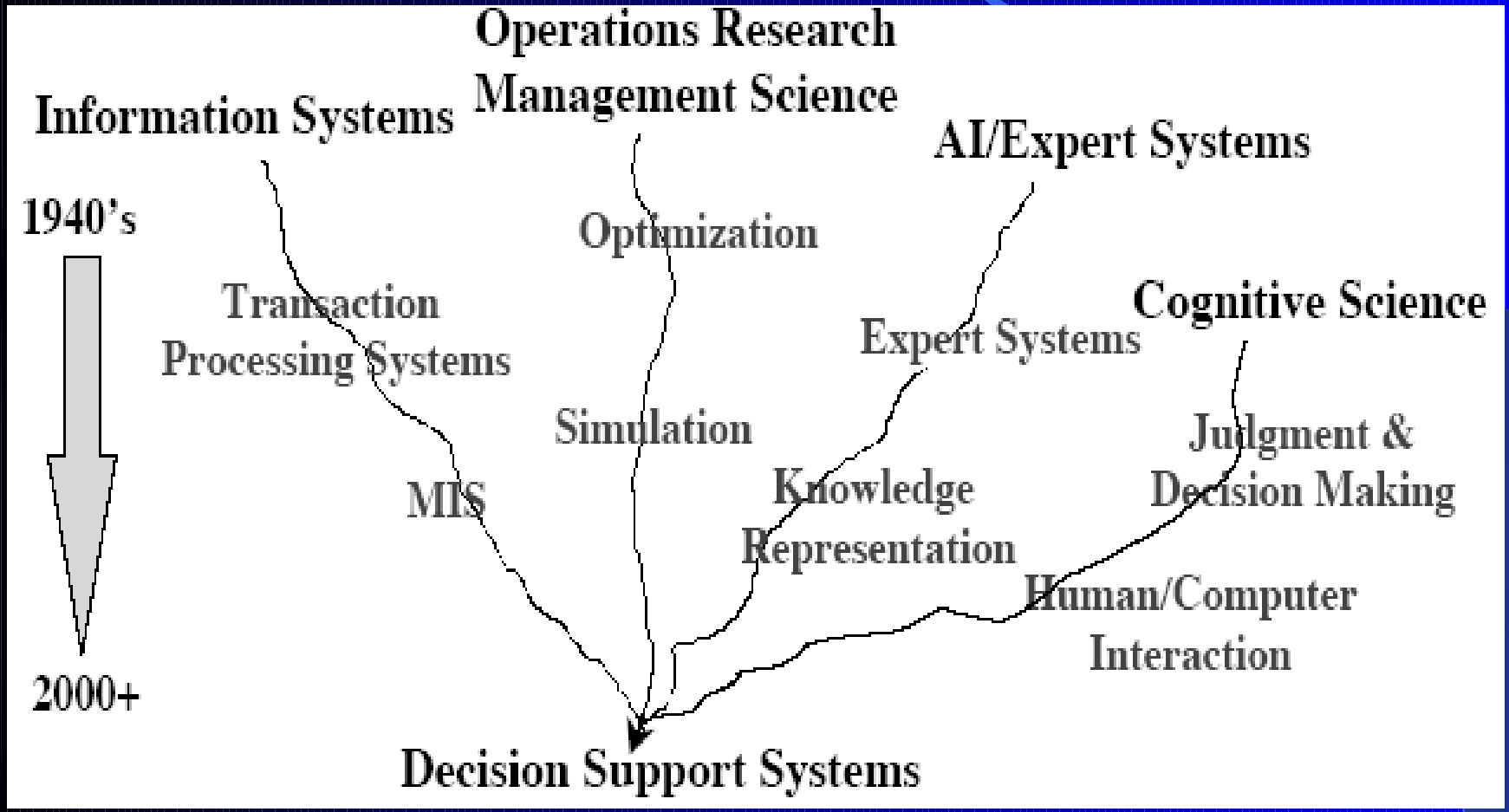
Introduction – Decision Support System (DSS)

- A **Decision Support System (DSS)** is an interactive computer-based system or subsystem intended to help decision makers use communications technologies, data, documents, knowledge and/or models to identify and solve problems, complete decision process tasks, and make decisions.
- It is clear that DSS belong to an environment with multidisciplinary foundations, including (but not exclusively):
  - Database research,
  - Artificial intelligence,
  - Human-computer interaction,
  - Simulation methods,
  - Software engineering, and
  - Telecommunications.

DSS

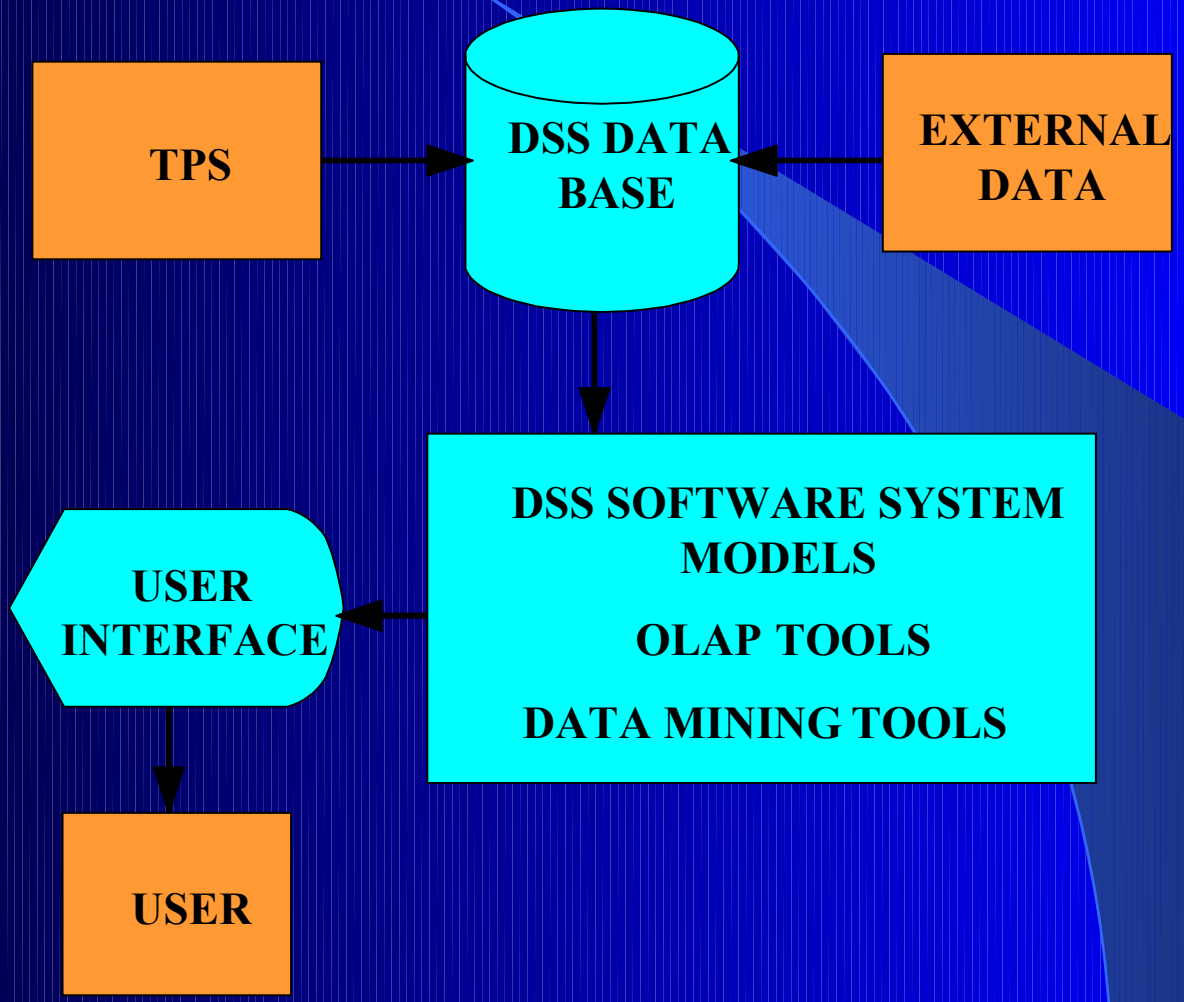
- A **Decision Support System (DSS)** is a computer-based information system that supports business or organizational decision-making activities.
- DSSs serve the management, operations, and planning levels of an organization (usually mid and higher management) and help to make decisions, which may be rapidly changing and not easily specified in advance (Unstructured and Semi-Structured decision problems).
- Decision support systems can be either fully computerized, human or a combination of both.

Historical Evolution of DSS



Typical DSS 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

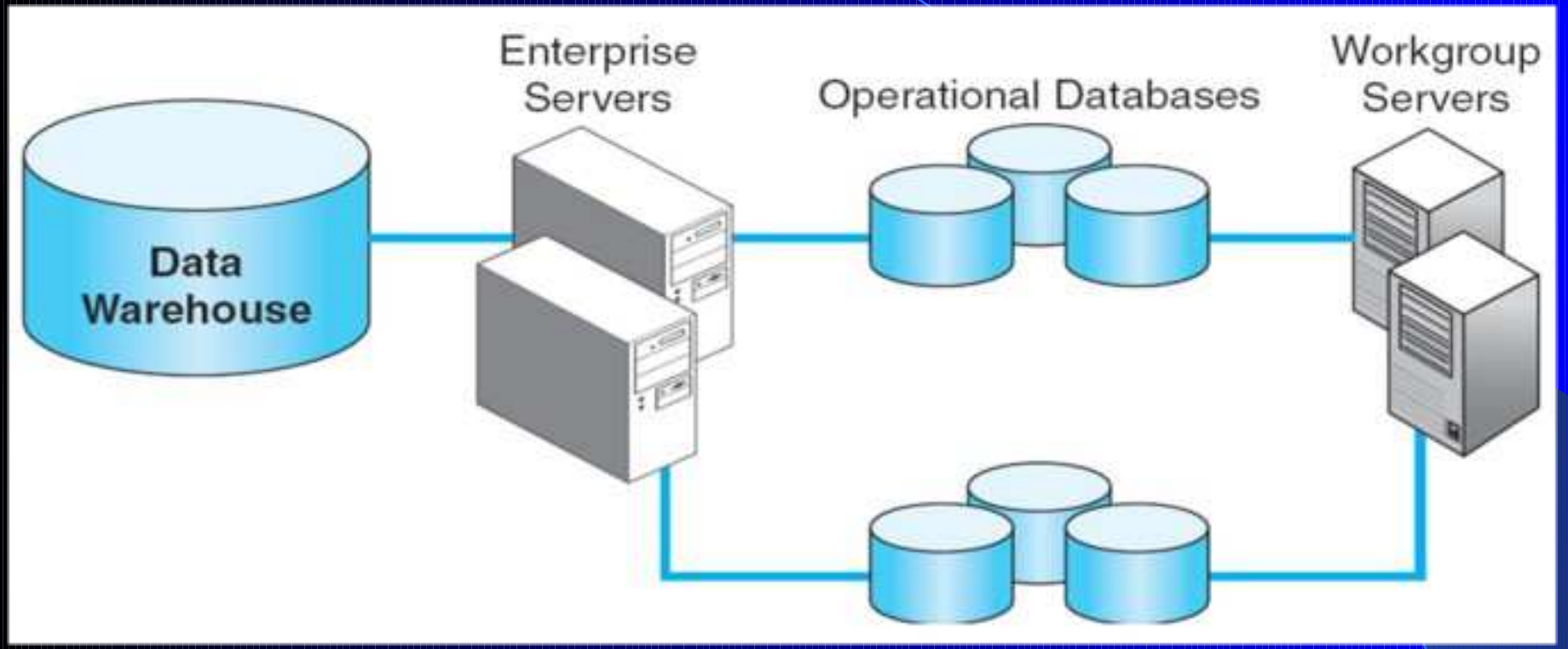


Why DSS?

- Increasing complexity of decisions
  - Technology
  - Information:
    - "Data, data everywhere, and not the time to think!"
  - Number and complexity of options
  - Pace of change
- Increasing availability of computerized support
  - Inexpensive high-powered computing
  - Better software
  - More efficient software development process
- Increasing usability of computers

Operational Databases

- Operational database management systems (also referred to as OLTP databases), are used to manage dynamic data in real-time.
- These types of databases allow you to do more than simply view archived data. Operational databases allows to modify that data (add, change or delete data), doing it in real-time.
- Since the early 90's, the operational database software market has been largely taken over by SQL engines.
- Today, the operational DBMS market (formerly OLTP) is evolving dramatically, with new, innovative entrants and incumbents supporting the growing use of unstructured data and NoSQL DBMS engines, as well as XML databases and NewSQL databases.
- Operational databases are increasingly supporting distributed database architecture that provides high availability and fault tolerance through replication and scale out ability.


Enterprise
Servers
Operational Databases
Workgroup
Servers
Data
Warehouse

Differences between the Databases and Data Warehouses

FEATURES	DATABASE	DATA WAREHOUSE
Characteristic	It is based on Operational Processing.	It is based on Informational Processing.
Data	It mainly stores the Current data which always guaranteed to be up-to-date.	It usually stores the Historical data whose accuracy is maintained over time.
Function	It is used for day-to-day operations.	It is used for long-term informational requirements and decision support.
User	The common users are clerk, DBA, database professional.	The common users are knowledge worker (e.g., manager, executive, analyst)
Unit of work	Its work consists of short and simple transaction.	The operations on it consists of complex queries..
Focus	The focus is on “Data IN”	The focus is on “Information OUT”
Orientation	The orientation is on Transaction.	The orientation is on Analysis.
DB design	The designing of database is ER based and application-oriented.	The designing is done using star/snowflake schema and its subject-oriented.
Summarization	The data is primitive and highly detailed.	The data is summarized and in consolidated form.
View	The view of the data is flat relational.	The view of the data is multidimensional.

I.T.S
GHAZIABAD

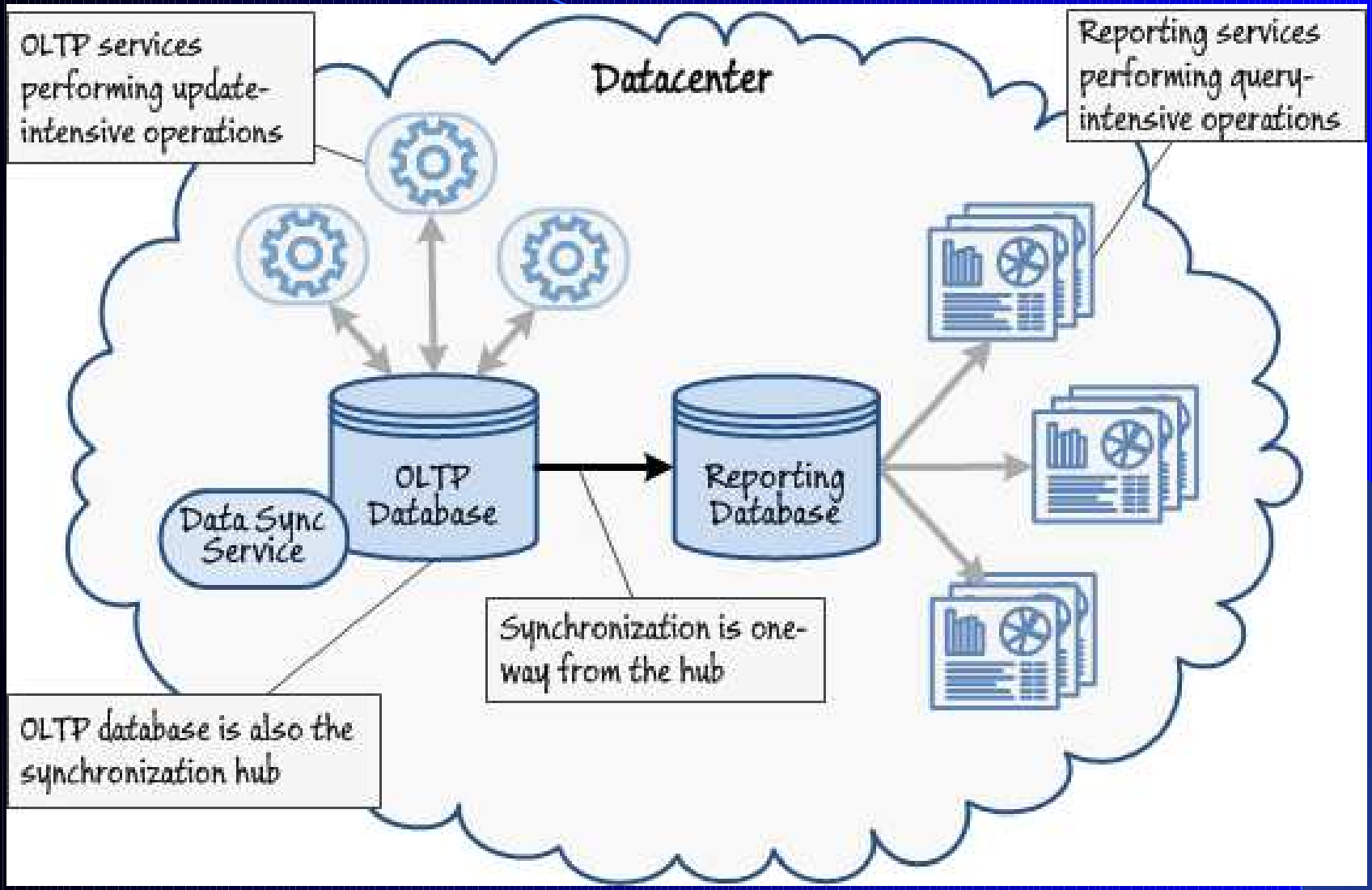
FEATURES	DATABASE	DATA WAREHOUSE
Function	It is used for day-to-day operations.	It is used for long-term informational requirements and decision support.
User	The common users are clerk, DBA, database professional.	The common users are knowledge worker (e.g., manager, executive, analyst)
Access	The most frequent type of access type is read/write.	It mostly use the read access for the stored data.
Operations	The main operation is index/hash on primary key.	For any operation it needs a lot of scans.
Number of records accessed	A few tens of records.	A bunch of millions of records.
Number of users	In order of thousands.	In the order of hundreds only.
DB size	100 MB to GB.	100 GB to TB.
Priority	High performance, high availability	High flexibility, end-user autonomy
Metric	To measure the efficiency, transaction throughput is measured.	To measure the efficiency, query throughput and response time is measured.

I.T.S
GHAZIABAD

DATA Warehousing - Introduction

A data warehouse is a subject-oriented, integrated, nonvolatile, time-variant collection of data in support of management's decisions.

- WH Inmon


Datacenter
OLTP services performing update-intensive operations
Reporting services performing query-intensive operations
OLTP Database
Reporting Database
Data Sync Service
Synchronization is one-way from the hub
OLTP database is also the synchronization hub

Data Warehouse Usage

- Three kinds of data warehouse applications
  - **Information processing**
    - supports querying, basic statistical analysis, and reporting using crosstabs, 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.
- Differences among the three tasks

Data Warehouse: Subject-Oriented

- Organized around major subjects, such as customer, product, sales.
- **Focusing on the modeling and analysis of data for decision makers, not on daily operations or transaction processing.**
- Provide a **simple and concise view around particular subject issues by excluding data that are not useful in the decision support process.**

Data Warehouse—Integrated

- Constructed by integrating multiple, heterogeneous data sources
  - relational databases, flat files, on-line transaction records
- Data cleaning and data integration techniques are applied.
  - Ensure consistency in naming conventions, encoding structures, attribute measures, etc. among different data sources
    - E.g., Hotel price: currency, tax, breakfast covered, etc.
  - When data is moved to the warehouse, it is converted.

Data Warehouse—Time Variant

- The time horizon 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, explicitly or implicitly
  - But the key of operational data may or may not contain "time element".

Data Warehouse—Non-Volatile

- 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
  - Requires only two operations in data accessing:
    - *initial loading of data* and *access of data*.

Data Warehouse vs. Heterogeneous DBMS

- Traditional heterogeneous DB integration:
  - Build wrappers/mediators on top of heterogeneous databases
  - Query driven approach
    - When a query is posed to a client site, a meta-dictionary is used to translate the query into queries appropriate for individual heterogeneous sites involved, and the results are integrated into a global answer set
    - Complex information filtering, compete for resources
- Data warehouse: update-driven, high performance
  - Information from heterogeneous sources is integrated in advance and stored in warehouses for direct query and analysis

Data Warehouse vs. Operational DBMS

- OLTP (on-line transaction processing)
  - Major task of traditional relational DBMS
  - Day-to-day operations: purchasing, inventory, banking, manufacturing, payroll, registration, accounting, etc.
- OLAP (on-line analytical processing)
  - Major task of data warehouse system
  - Data analysis and decision making
- Distinct features (OLTP vs. OLAP):
  - User and system orientation: customer vs. market
  - Data contents: current, detailed vs. historical, consolidated
  - Database design: ER + application vs. star + subject
  - View: current, local vs. evolutionary, integrated
  - Access patterns: update vs. read-only but complex queries

I.T.S GHAZIABAD

OLTP vs. OLAP

	OLTP	OLAP
users	clerk, IT professional	knowledge worker
function	day to day operations	decision support
DB design	application-oriented	subject-oriented
data	current, up-to-date detailed, flat relational isolated	historical, summarized, multidimensional integrated, consolidated
usage	repetitive	ad-hoc
access	read/write index/hash on prim. key	lots of scans
unit of work	short, simple transaction	complex query
# records accessed	tens	millions
#users	thousands	hundreds
DB size	100MB-GB	100GB-TB
metric	transaction throughput	query throughput, response

Why Separate Data Warehouse?

- High performance for both systems
  - 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:** DS 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 reconciled

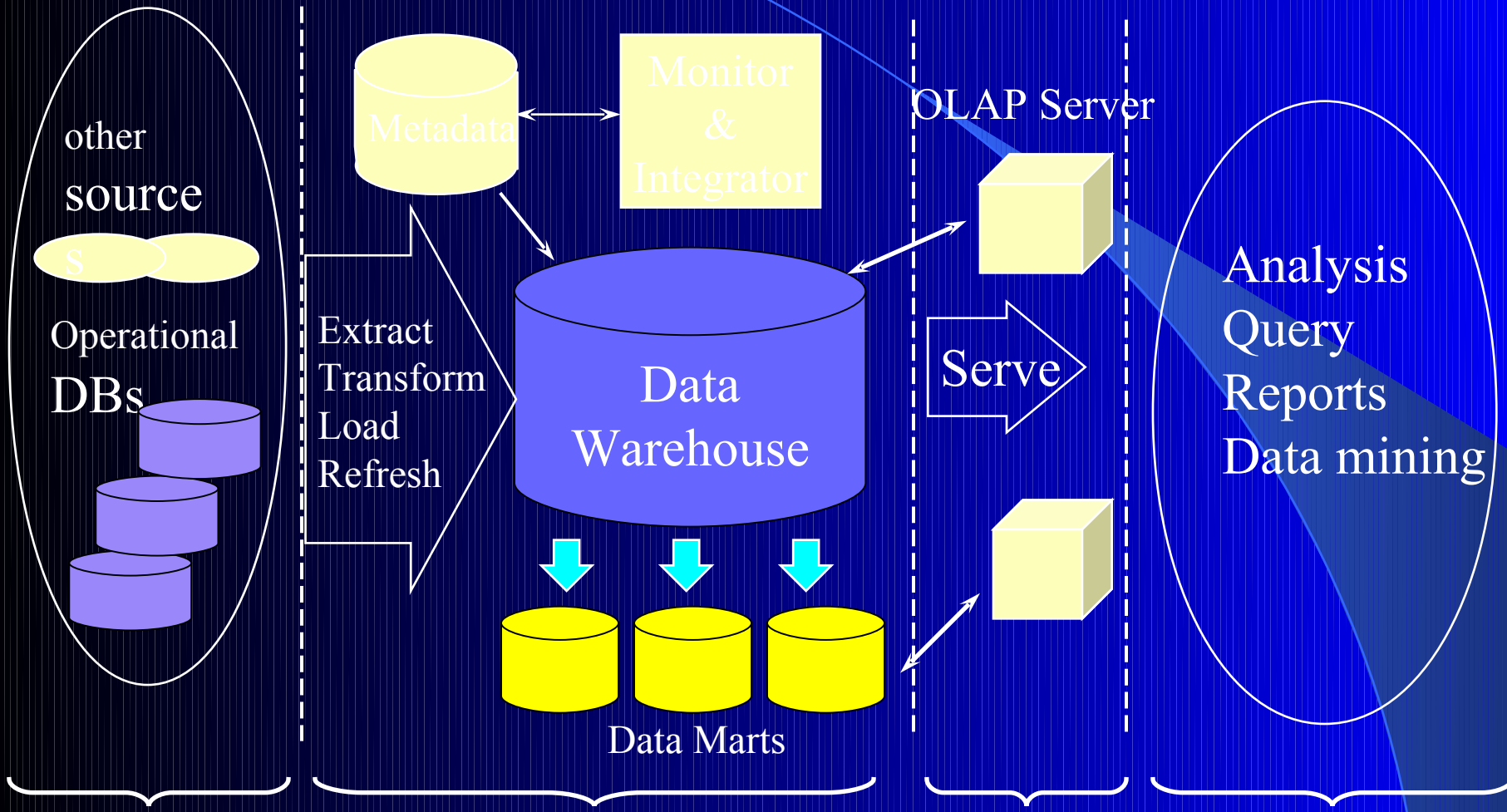
I.T.S
GHAZIABAD

Data Mart

I.T.S
GHAZIABAD

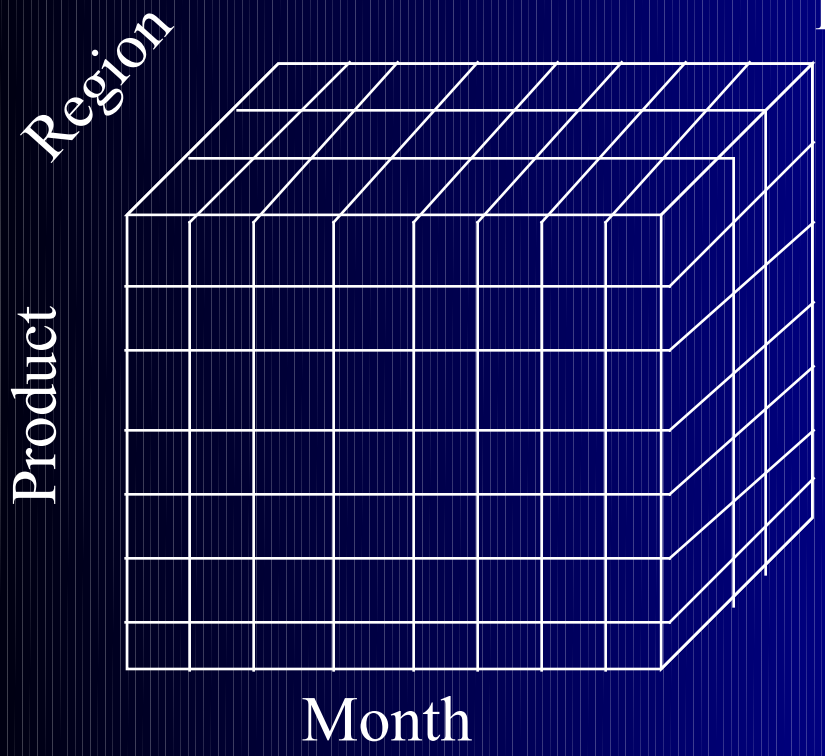
Concept of Data-Warehousing

Multi-Tiered Architecture

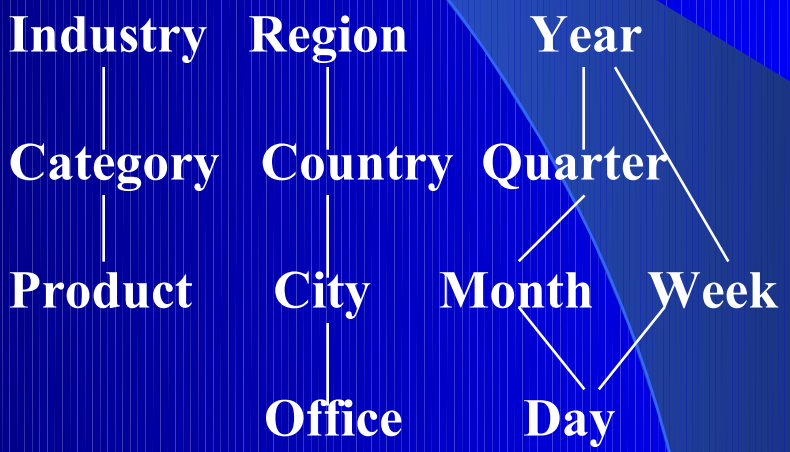


Multi Dimensional Database Structures

- Sales volume as a function of product, month, and region



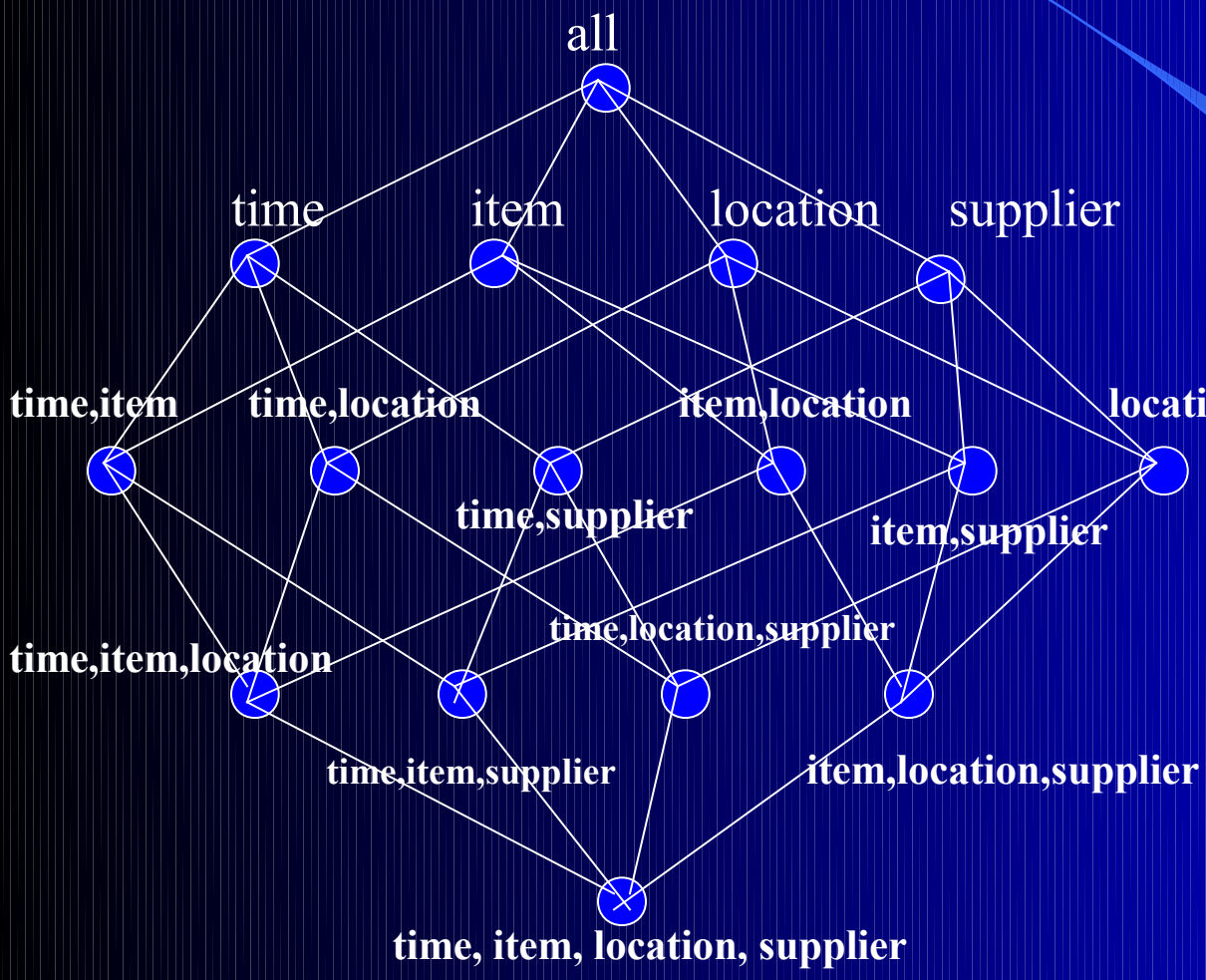
Dimensions: Product, Location, Time
Hierarchical summarization paths



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
- In data warehousing literature, an n-D base cube is called a **base cuboid**. The top most 0-D cuboid, which holds the highest-level of summarization, is called the **apex cuboid**. The lattice of cuboids forms a **data cube**.

Cube: A Lattice of Cuboids



0-D(apex) cuboid

1-D cuboids

2-D cuboids

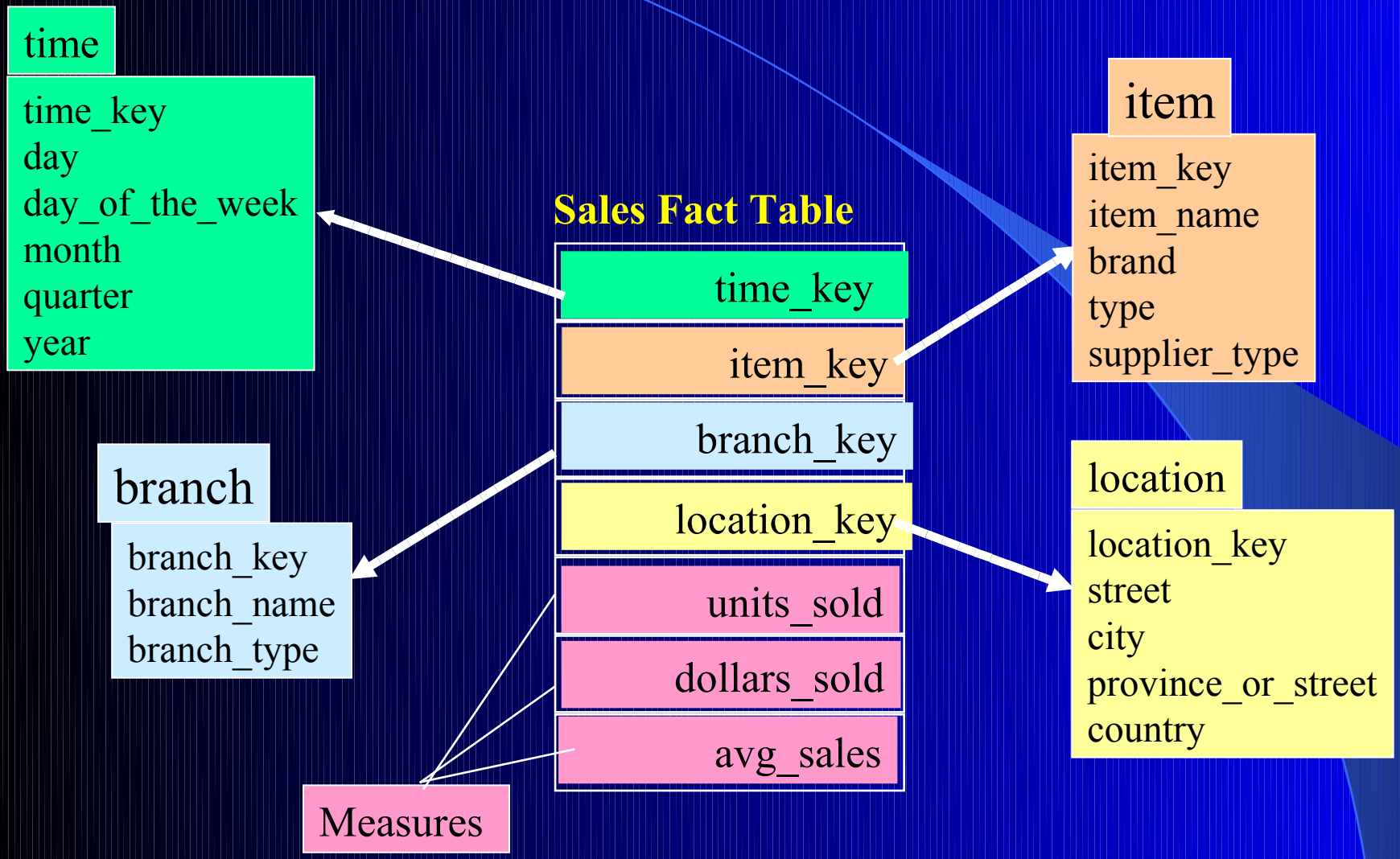
3-D cuboids

4-D(base) cuboid

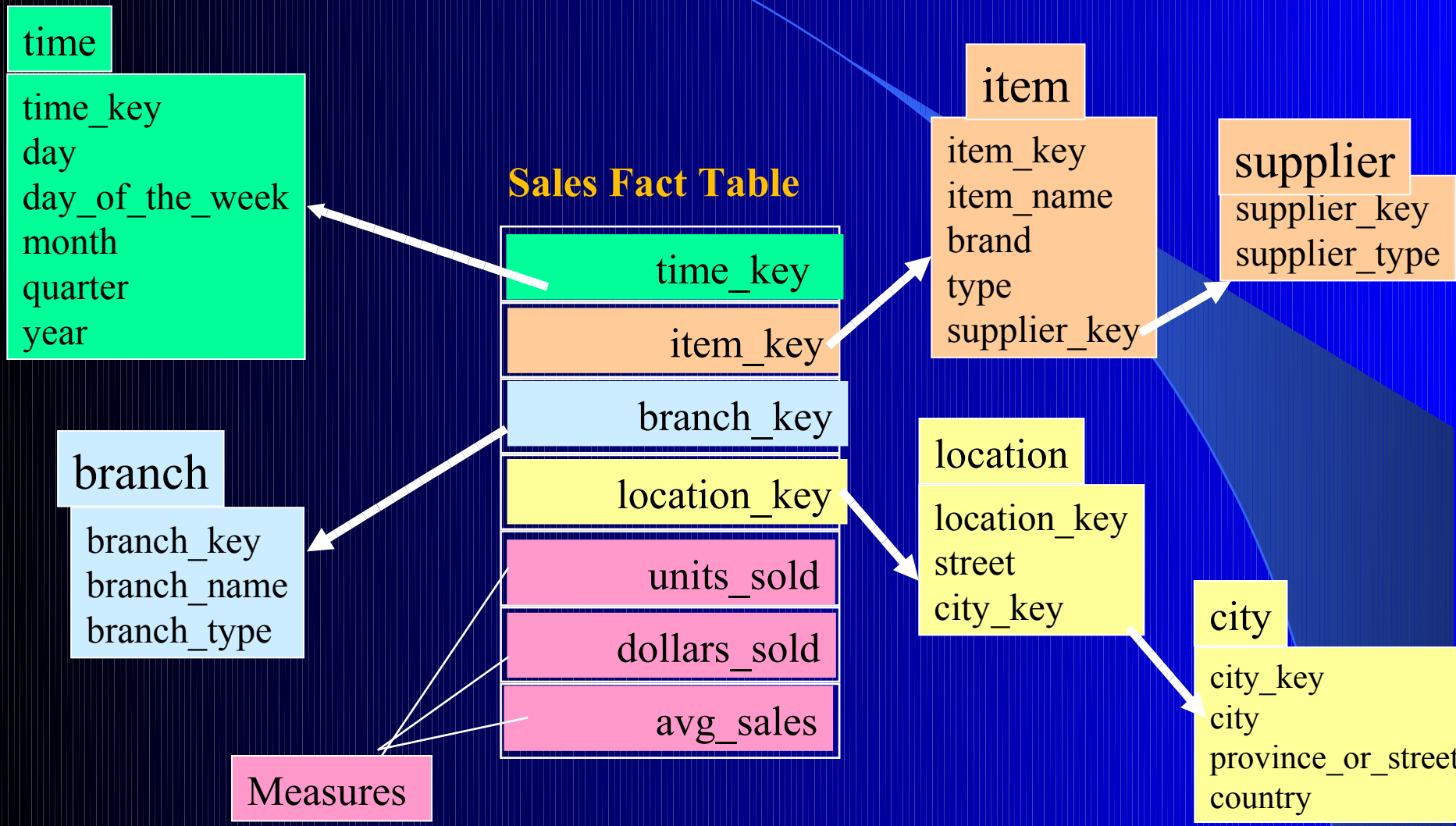
Conceptual Modeling of Data Warehouses

- 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, forming a shape similar to snowflake
  - **Fact constellations: Multiple fact tables share dimension tables**, viewed as a collection of stars, therefore called galaxy schema or fact constellation

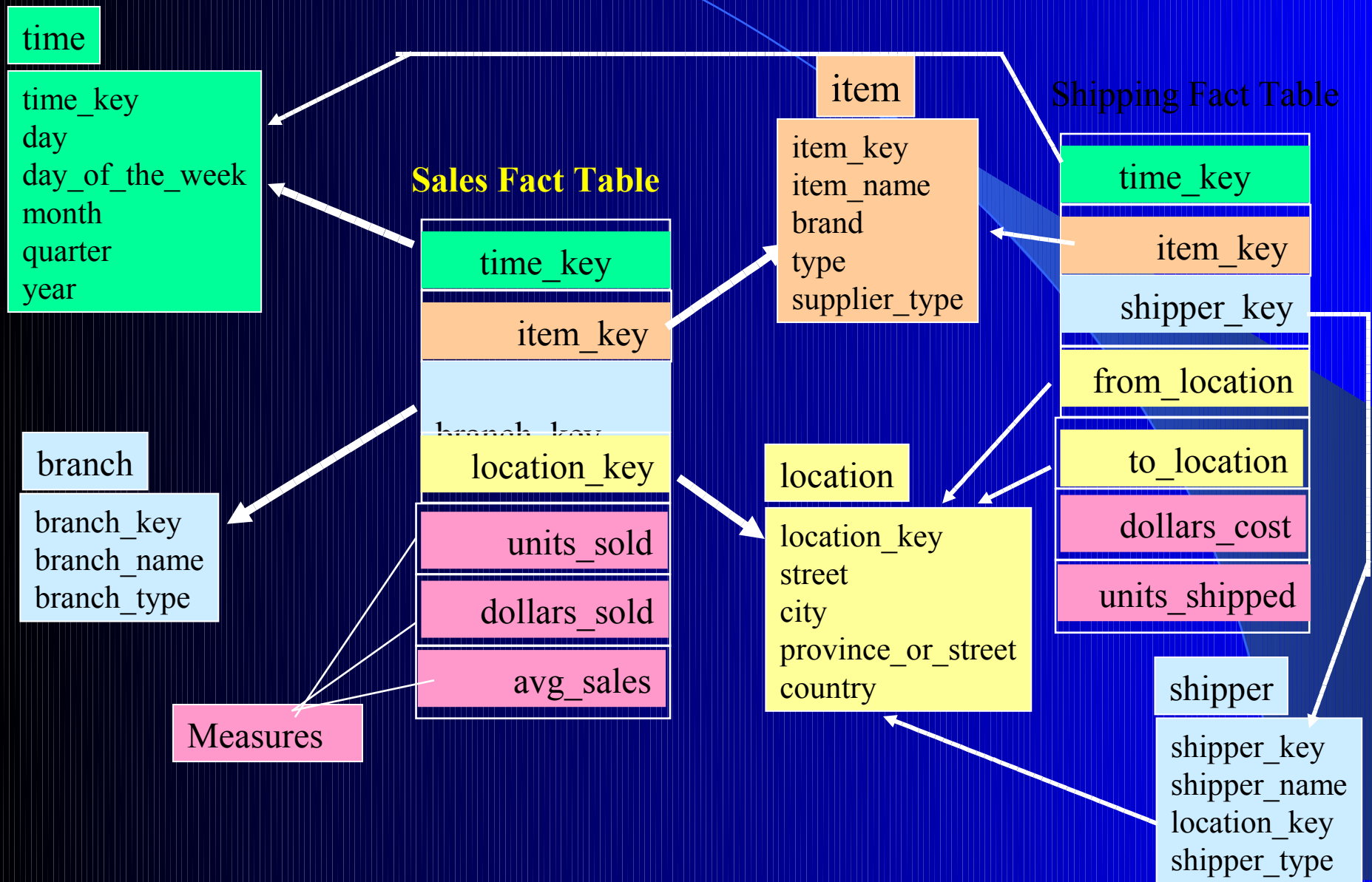
Example of Star Schema



Example of Snowflake Schema



Example of Fact Constellation



Client/Server Computing Model & Data Warehousing

- The fundamental characteristic of client/server computing is distribution of computing resources (e.g. data, compute power) across different computers.
- The idea is to divide applications into logical segments (tasks) so that they are then performed on platforms most appropriate.
- A **client/server database system** increases processing power by separating the database management system from the application; the client as the front-end system handling the user interface and the server as the back-end system accessing the database, which cooperate to run an application.

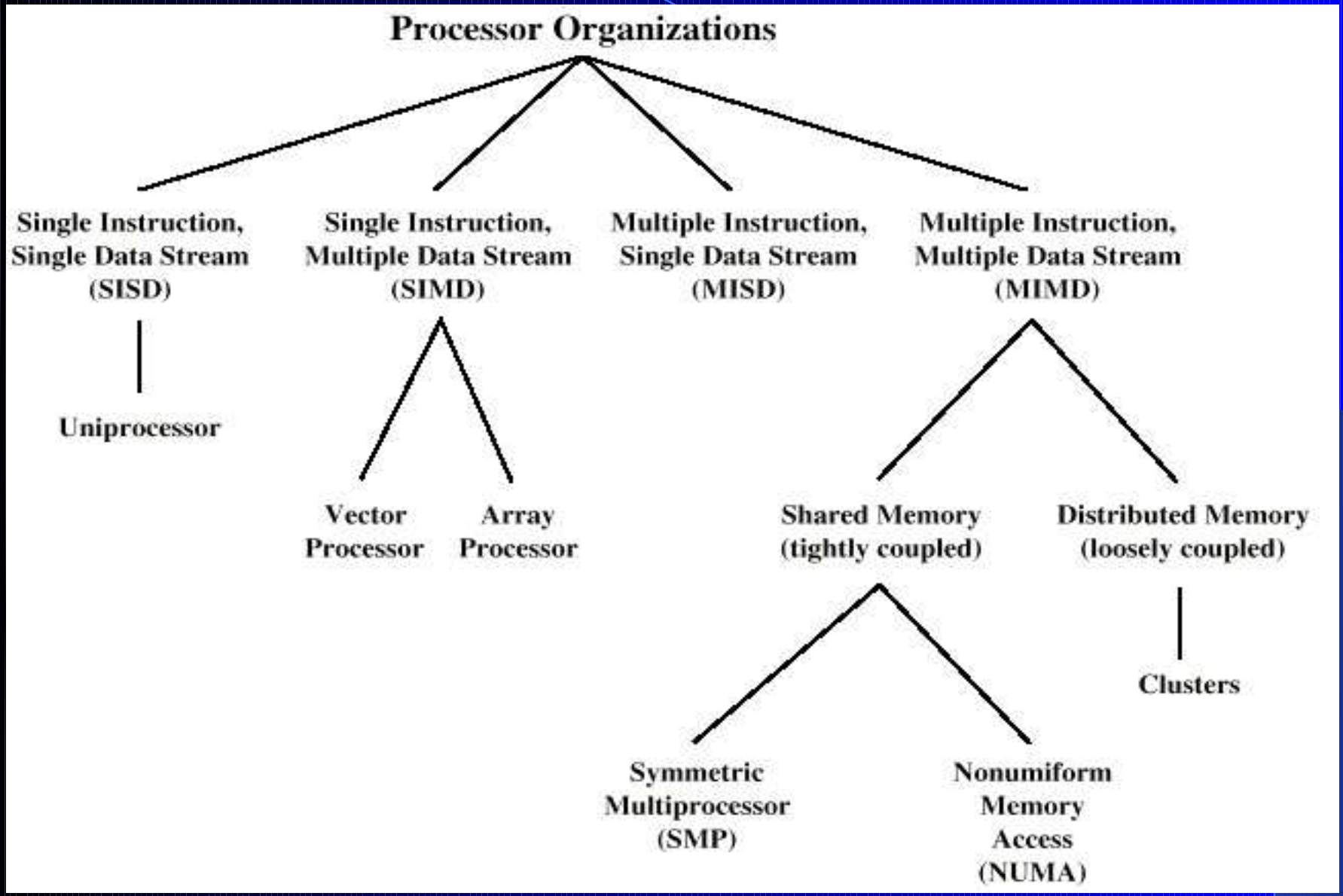
Contd....

- Data Warehousing is a continual process which enables a corporation to assemble operational and other data from a variety of internal and external sources, and transform that data into consistent, high-quality, business information, distribute that information to the points of maximum value within the organizations, and provide easy, flexible and fast access for busy non-technical users.

Reasons for using client/server

- Exploitation of centralised computing power /data capacity
- Scalability
- Performance
- Flexibility (in order to adjust to changing demands)
- GUI on desktop
- Protection of investment, strategic software, strategic data
- Client/server provides an integrated solution.

Parallel Processors & Cluster Systems



Loosely Coupled - Clusters

- Collection of independent whole uni-processors or SMPs
  - Usually called nodes
- Interconnected to form a cluster
- Working together as unified resource
  - Illusion of being one machine
- Communication via fixed path or network connections

Cluster Benefits

- Absolute scalability
- Incremental scalability
- High availability
- Superior price/performance

Distributed DBMS implementations