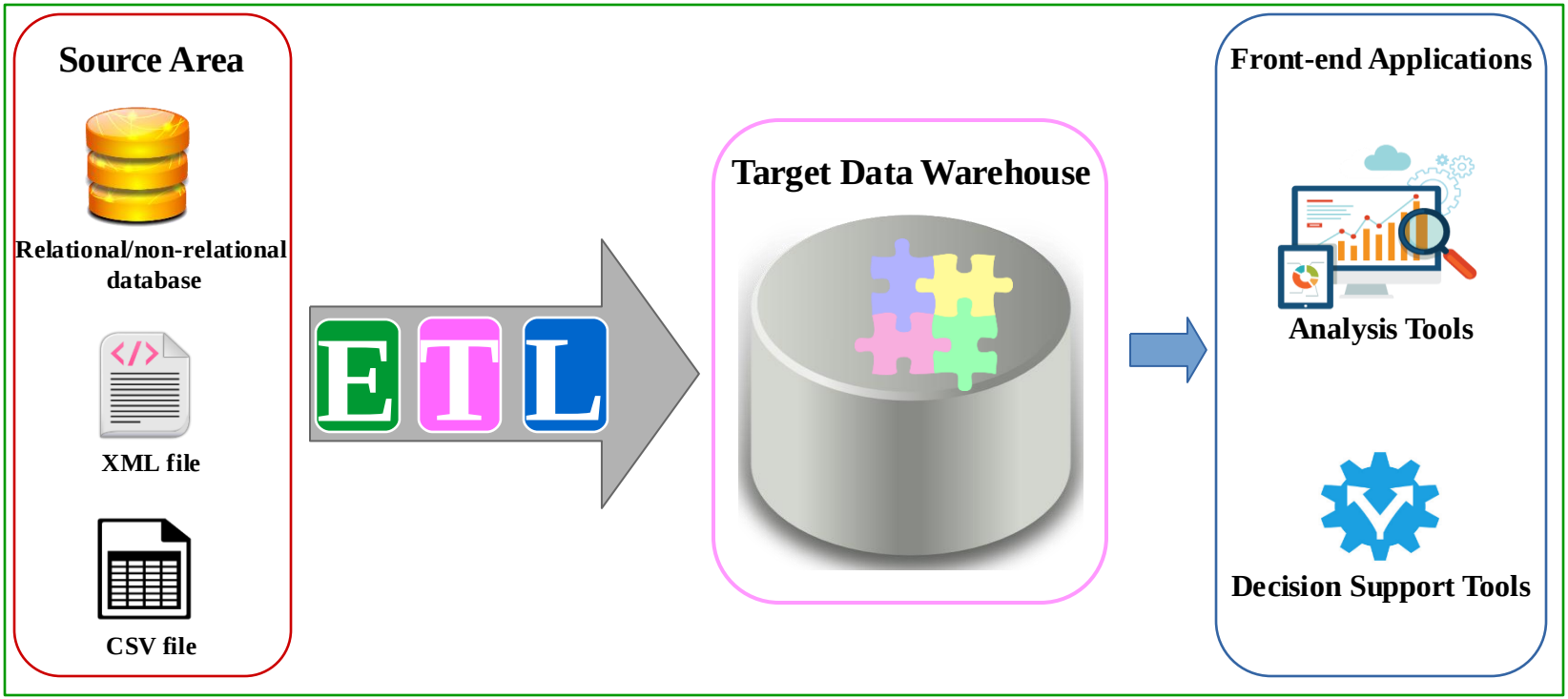
Data Warehouse Systems

Sean Kang

Data Warehouse System



Extract-Transform-Load Process

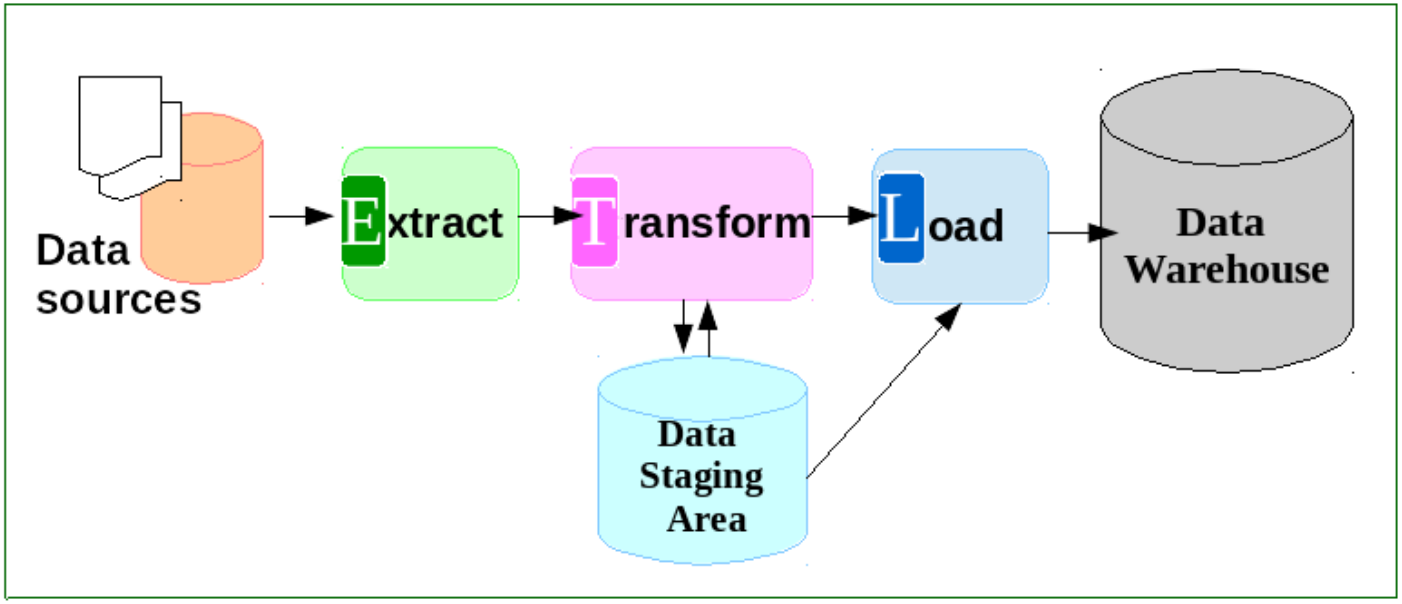


Table Mapping Examples

- One-to-one table mapping

Source Table	Source Attribute	Target Table	Target Attribute	Selection Condition
Address	Address_key	Location	Location_id	Transform all the new addresses (Year>2000)

- Many-to-one table mapping

Source Table	Source Attribute	Target Table	Target Attribute	Selection Condition
Patient, Address	Address_key, Address_key	Person	Location_id	Transform all the current patients with their new addresses (Year>2000)

LEFT JOIN

Attribute Mapping Examples

- One-to-one attribute mapping

Source Table	Source Attribute	Target Table	Target Attribute	Selection Condition
Address	Address_key	Location	Location_id	Transform all the new addresses (Year>2000)

- Many-to-one attribute mapping

Source Table	Source Attribute	Target Table	Target Attribute	Selection Condition
Patient	Day_of_birth, Month_of_birth, Year_of_birth	Person	Date_of_birth	Transform all the current patients

Data Warehouse Features (Separation of Purposes)

- A decision support database that is maintained separately from the organization's operational database
- Supports information processing by providing a solid platform of consolidated, historical data for analysis
- Focuses on the modeling and analysis of data for decision makers, not on daily operations or transaction processing

Data Warehouse vs. Operational DBMS

Major task of traditional relational DBMS:

- OLTP (on-line transaction processing)
- Day-to-day operations, such as purchasing, inventory, banking, manufacturing, payroll, registration, and accounting

Major task of data warehouse system:

- OLAP (on-line analytical processing)
- Data analysis and decision making

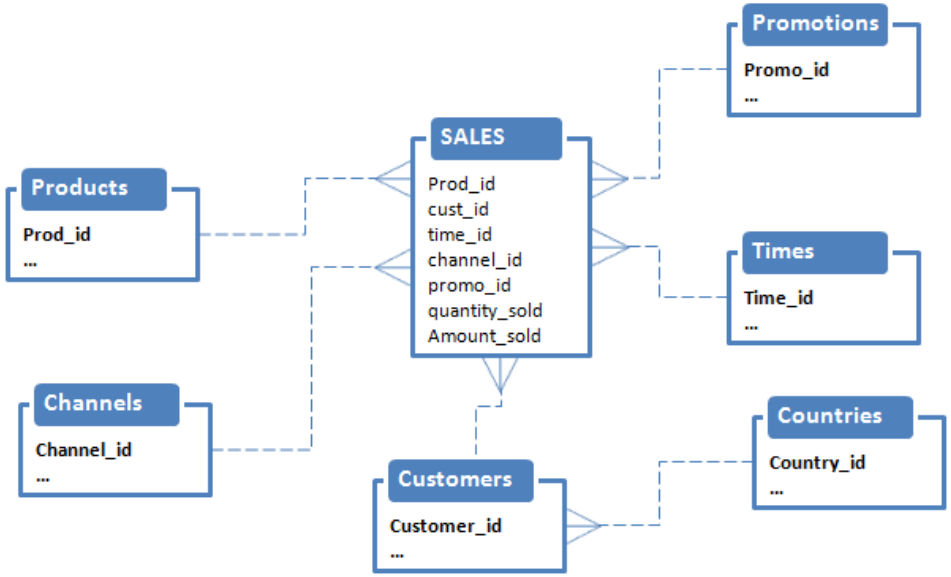
OLTP vs. OLAP

	OLTP	OLAP
users	customer	knowledge worker
function	day to day operations	decision support
DB design	relational model	multidimensional model
data	current, up-to-date detailed	historical, summarized, integrated
access	read/write	read lots of scans
unit of work	short, simple transaction	complex query
# records accessed	tens	millions
#users	thousands	hundreds
DB size	100MB-GB	100GB-TB

Multidimensional Data Model

Structures optimized for end-user queries that include:

- Dimension table: contains descriptive attributes
- Fact table: contains measurements of a business process and keys to each of the related dimension tables



Multidimensional Data Model

Star schema:

- A fact table in the middle connected to a set of dimension tables

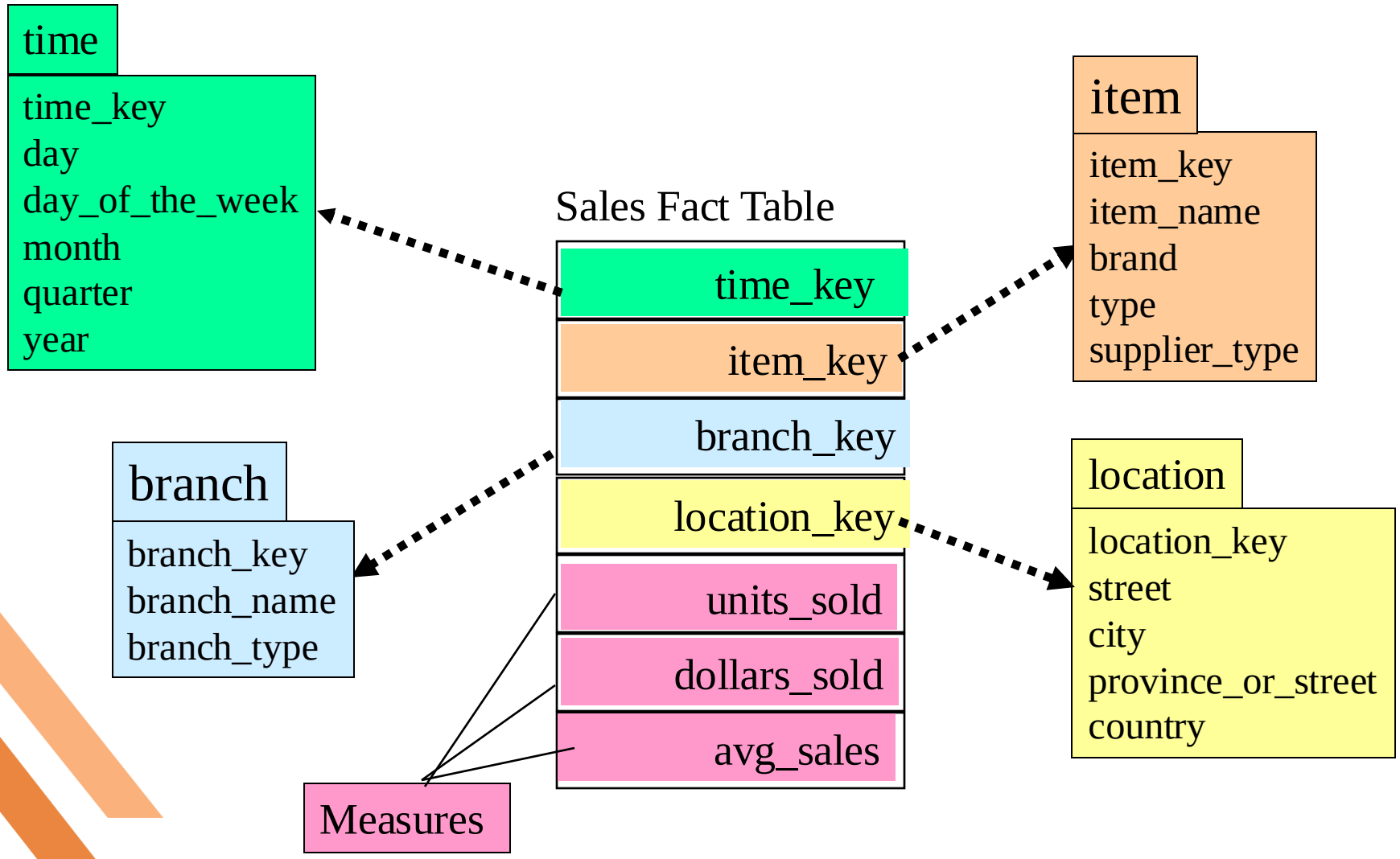
Snowflake schema (diagram on next slide)

- Dimensions are split into more than one dimension tables
- Star schema is a special case of the snowflake schema with a single level hierarchy

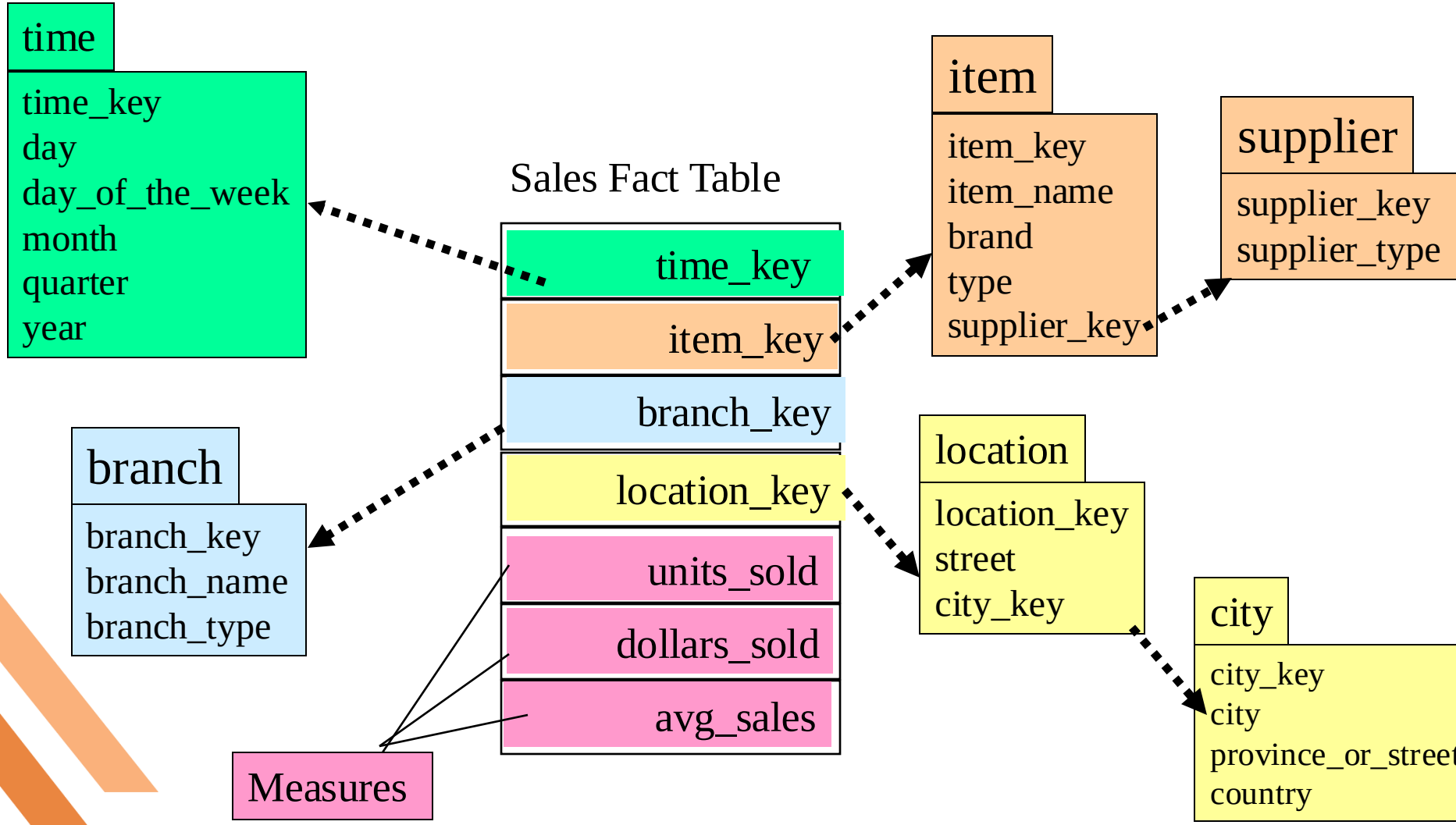
Fact constellations (diagram on next slide)

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

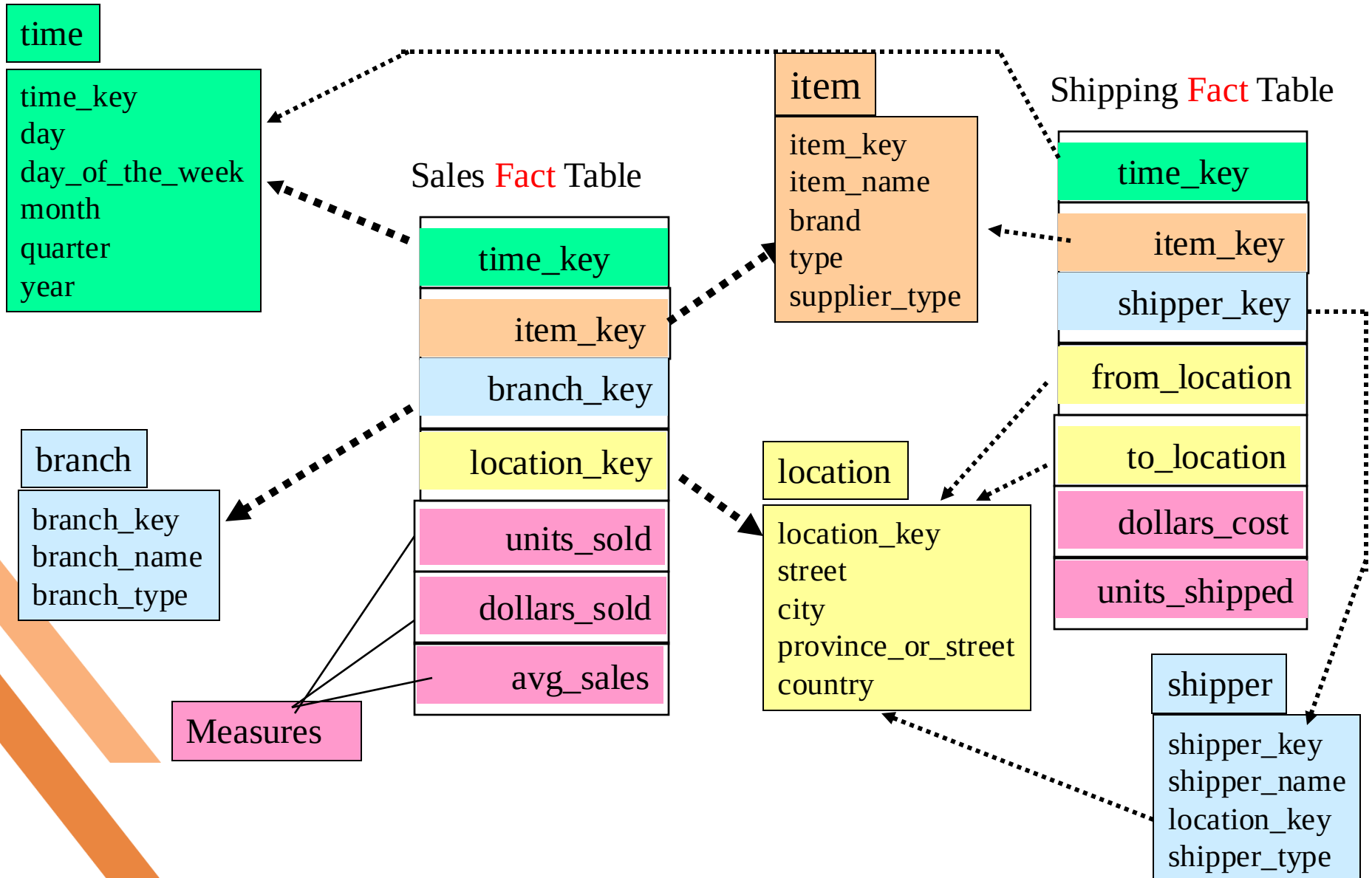
Example of Star Schema



Example of Snowflake Schema



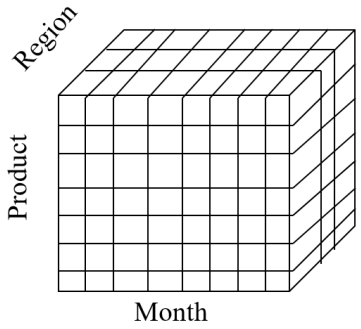
Example of Fact Constellation



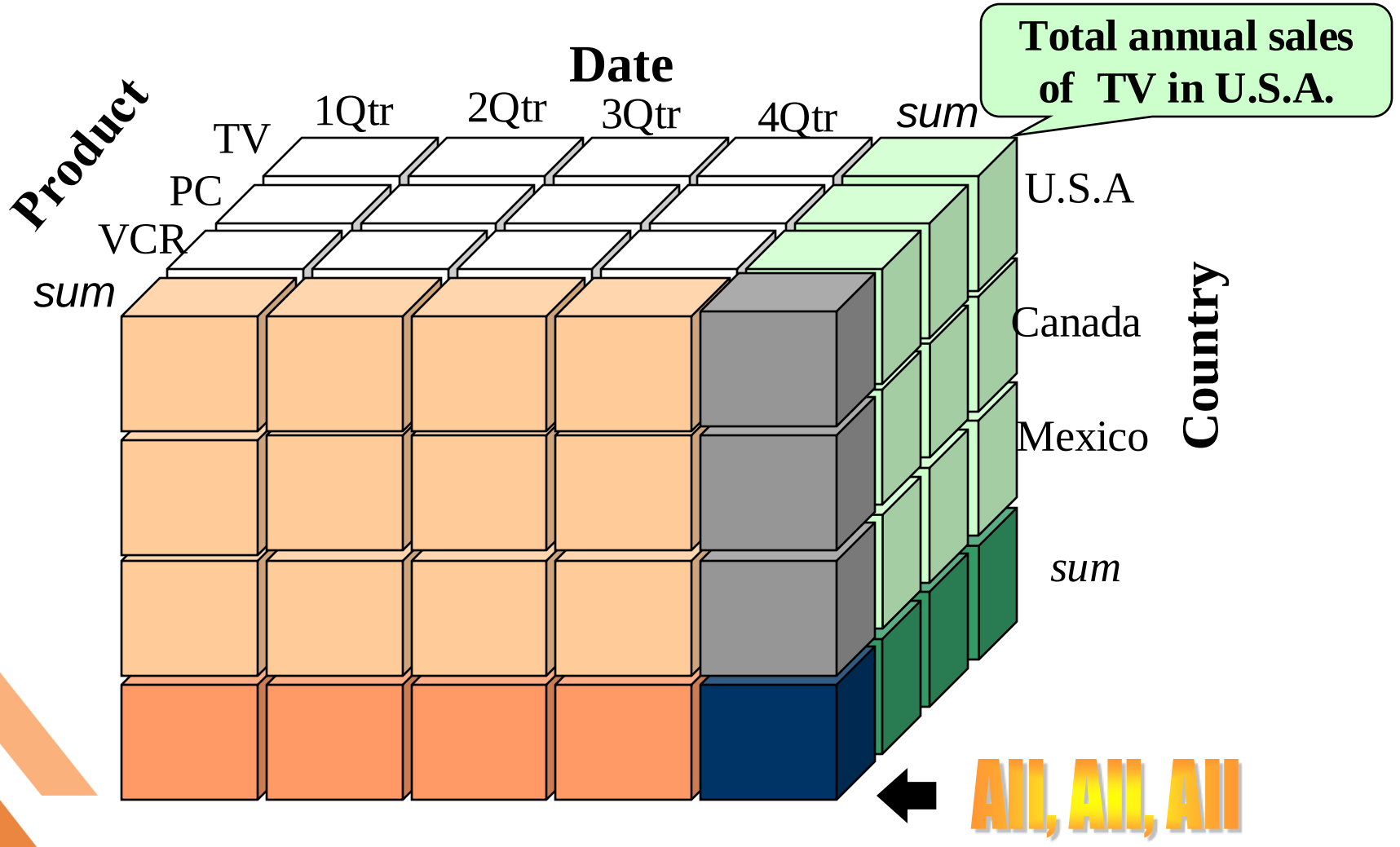
OLAP Using Multidimensional Data

- Enable users to access a wide variety of views of data for multidimensional analysis
- Unlike traditional relational reports that represent data in two-dimensional row and column format, represent their aggregated data in a multi-dimensional structure called cube
- Supports complicated queries involving facts to be measured across different dimensions

Example: Sales volume as a function of product, month, and region



A Sample Data Cube



Data Warehouse Usage

Information processing

- Supports querying, basic statistical analysis, and reporting using tables, charts and graphs

Analytical processing

- Multidimensional analysis of data warehouse data using OLAP operations

Data mining

- Knowledge discovery from hidden patterns
- Supports associations, constructing analytical models, performing classification and prediction, and presenting the mining results using visualization tools

Summary

- Enterprises use data warehouses to accumulate data from multiple sources for data analysis and research
- ETL process involves extracting data from source databases, transforming it into a form suitable for research and analysis, and loading it into a data warehouse
- DBMS is tuned for OLTP: access methods, indexing, concurrency control, recovery
- Data warehouse is tuned for OLAP: complex OLAP queries, multidimensional view, consolidation

Case Study Example: Transportation

- Your Trip to NYC and back
- One or More Stop-overs
  - Each Flight contains
    - Aircraft Tail ID, Departure, Destination, Planned Start time, Planned End Time,
    - Aircraft Crew
    - Aircraft Maintenance and Operations Status
    - Flight Path, Runway used on take off, landing,
    - Capacity at take off, Capacity ramp-up from 60 days
  - Your Reservation
    - Passenger Name, age, gender, contact info, purchase date, purchase source
    - Number of days before departure
    - Class level, mileage or reward accumulation

Tables

- How many Fact Tables Involved in this Case Study
- Multi-level Fact Tables
  - Aircraft Flight
  - Passenger Reservation
  - Aircraft
- How many Dimension Tables Involved
  - Country, City, Airport
  - Aircraft, Fuel
  - Currency, Reward Points
  - Timezone, Country-code,
  - Passenger Type, Seat-Class,

Analysis Questions

- How often do passengers upgrade? (is there sufficient data)
- How % of customers earn and redeem their rewards within a year? (is there sufficient data)
- Do they respond to fare promotions? (does the reservation data contain promotion codes)
- Do customer take flights with overnight stop-over?
- What proportion of the flight contains platinum passengers?
- How many days before departure did most passengers reserve their tickets?
- How many seats were empty on the flight?
- Was the flight used as a multi-leg of a journey ?

Sample Table