RISTECH
RESEARCH AND INNOVATION
IN SCIENCE ENGINEERING AND TECHNOLOGY

Data Warehousing

Ali Saeed Khan

BI (Business Intelligence)

- Business intelligence is the process of integrating enterprise-wide data into a single data store from which end users can run ad hoc queries and reports to analyze the existing data.
- In other words, the goal of BI is to keep data that can be accessed by users who make their business decisions on the basis of the analysis.
- These systems are often called *analytic* or *informative* systems because, by accessing data, users get the necessary information for making better business decisions.

OLTP Systems

- Online Transaction Processing Systems
- An atm transaction
- Relational and other type of data base routine operations

BI Goals

- The goals of BI systems are different from the goals of OLTP systems.
- The following is a query that is typical for a BI system:

"What is the best-selling product category for each sales region in the third quarter of the year 2011?"

- The most important properties of a BI system are as follows:
  - Periodic write operations (load) with queries based on a huge number of rows
  - Small number of users
  - Large size of data stored in a database

BI (Business Intelligence)

- Other than loading data at regular intervals (usually daily), BI systems are mostly read-only systems.
- In contrast to databases in OLTP systems that store only current data, BI systems must also track historical data.

OLTP vs. BI (OLAP)

RISTECH
RESEARCH AND INNOVATION
IN SCIENCE ENGINEERING AND TECHNOLOGY

	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

OLTP vs. OLAP

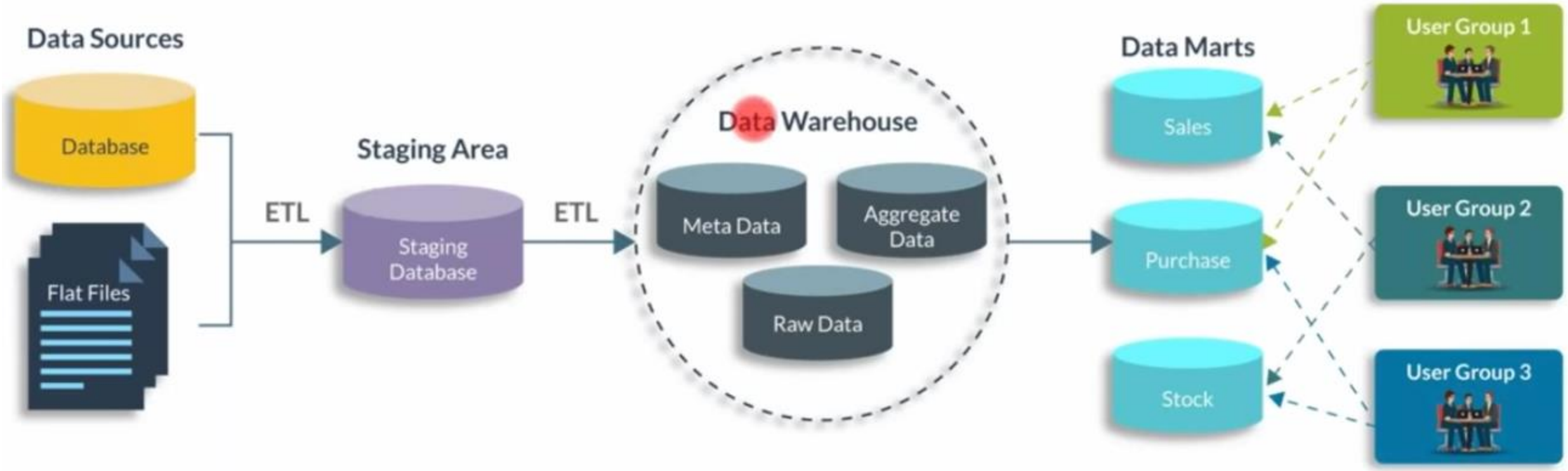
- A supermarket server which records every single product purchased at that market
- A bank server which records every time a transaction is made for a particular account
- Bank Manager wants to know how many customers are utilizing are utilizing the ATM of his branch. Based on this he may take call whether to continue with ATM or relocate it

Data Warehouse (DWH) ???

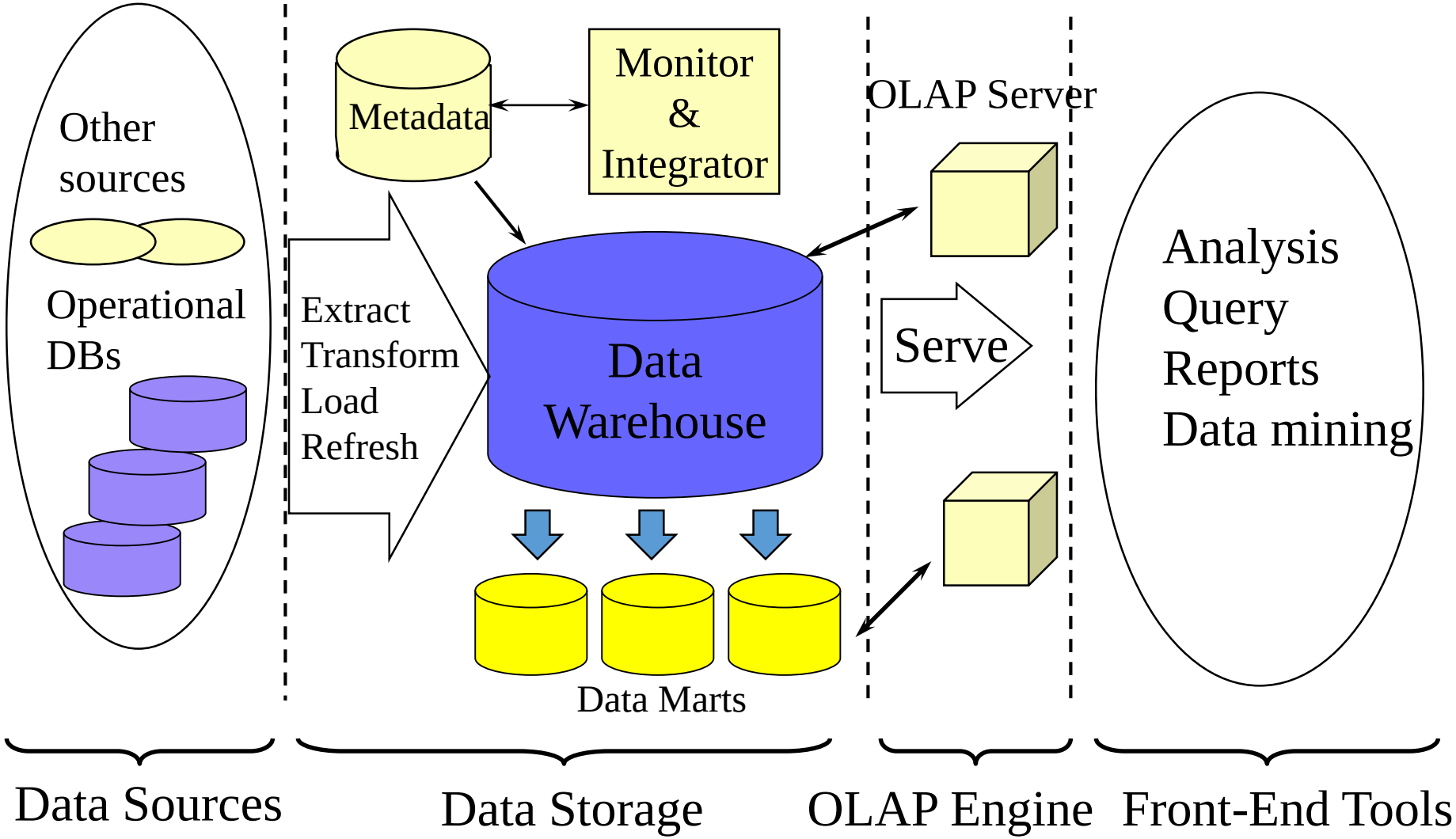
RISTECH
RESEARCH AND INNOVATION
IN SCIENCE ENGINEERING AND TECHNOLOGY

- Defined in many different ways, but not rigorously.
  - A decision support database that is maintained separately from the organization's operational database
  - Support information processing by providing a solid platform of consolidated, historical data for analysis.
- "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
- Data warehousing:
  - The process of constructing and using data warehouses

DWH Architecture



DWH: A Multi-Tiered Architecture



ETL (Extract, Transform, Load)

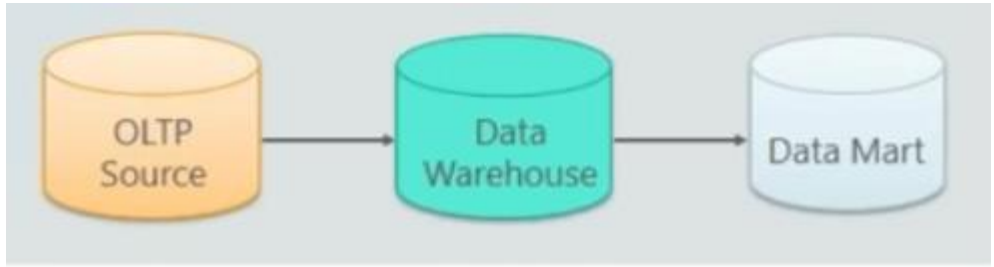
Data Mart

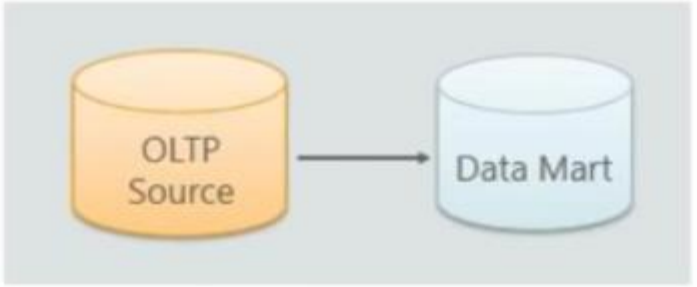
- Data mart is smaller version of the Data warehouse which deals with a single subject
- Data marts are focused on one area. Hence, they draw data from a limited number of sources

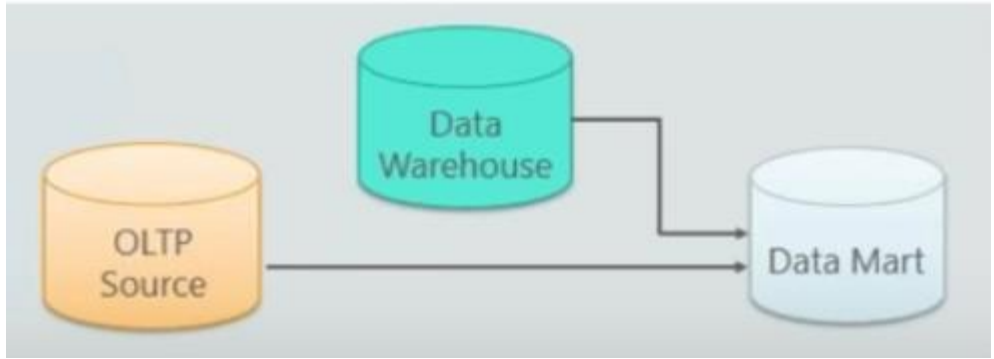
Data Warehouse	Data Marts
Enterprise wide data	Department wide data
Multiple subject areas	Single subject area
Multiple data sources	Limited data sources
Occupies large memory	Occupies limited memory
Longer time to implement	Shorter time to implement

Types of Data Mart

- Dependent
- Independent
- Hybrid







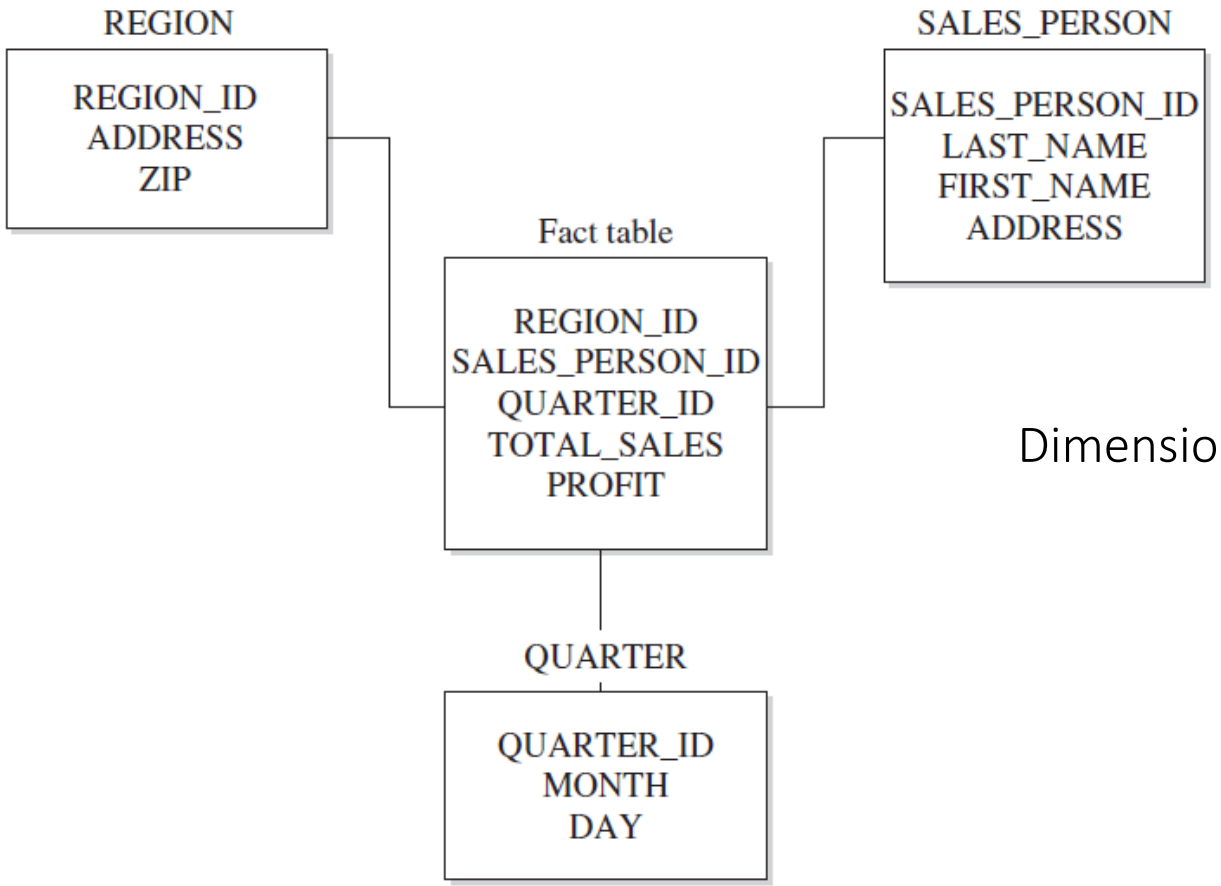
DWH Design

- Relational databases and data warehouses have a lot of differences that require different design methods.
- Relational databases are designed using the well known entity-relationship (ER) model, while the dimensional model is used for the design of data warehouses and data marts.
- BI processes are based on queries that operate on a huge amount of data and are neither simple nor short.
  - Therefore, the highly normalized tables do not suit the design of data warehouses, because
  - the goal of BI systems is significantly different: there are few concurrent transactions, and each transaction accesses a very large number of records.

Dimensional Model

- In dimensional modeling, every particular model is composed of one table (***fact table***) that stores measures and several other tables (***dimension***) that describe dimensions. The former is called the
- Examples of data stored in a fact table include inventory sales and expenditures.
  - Dimension tables usually include time, account, product, and employee data.

Dimensional Model



Dimensional Model: Star Schema

Dimensional Model

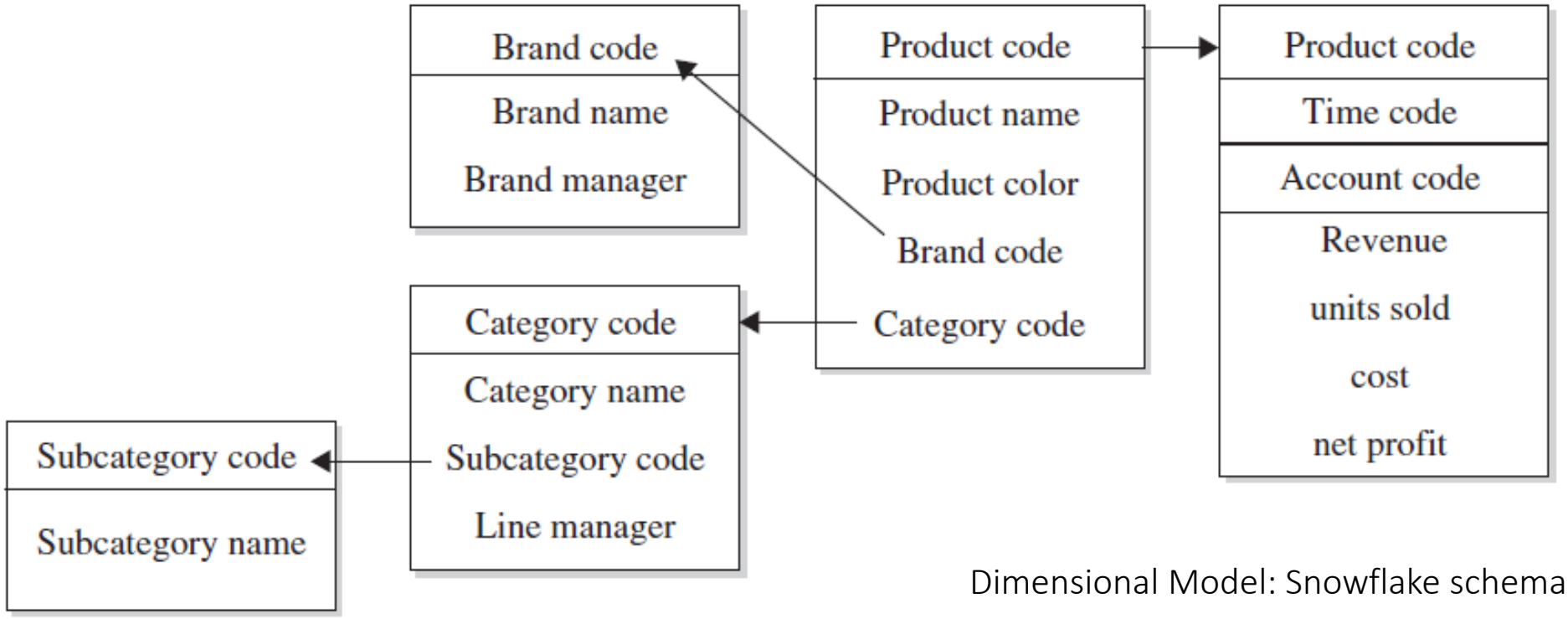
- Another difference in the nature of data in a fact table and the corresponding dimension tables is that most non key columns in a fact table are numeric and additive, because such data can be used to execute necessary calculations.
  - For example, columns like **Units_of_ product_sold**, **Total_sales**, **Profit**, or **Dollars_cost** are typical columns in the fact table.
- Numerical columns of the fact table that do not build the primary key of the table are called ***measures***.

Dimensional Model

RISTECH
RESEARCH AND INNOVATION
IN SCIENCE ENGINEERING AND TECHNOLOGY

- Columns of dimension tables are usually highly *denormalized*, which means that a lot of columns depend on each other.
- The denormalized structure of dimension tables has one important purpose: all columns of such a table are used as column headers in reports.
- If the denormalization of data in a dimension table is not desirable, a dimension table can be decomposed into several sub-tables.
- This is usually necessary when columns of a dimension table build hierarchies.
  - (For example, the **product** dimension could have columns such as **Product_id**, **Category_id**, and **Subcategory_id** that build three hierarchies, with the primary key, **Product_id**, as the root.)

Dimensional Model



Dimensional Model: Snowflake schema

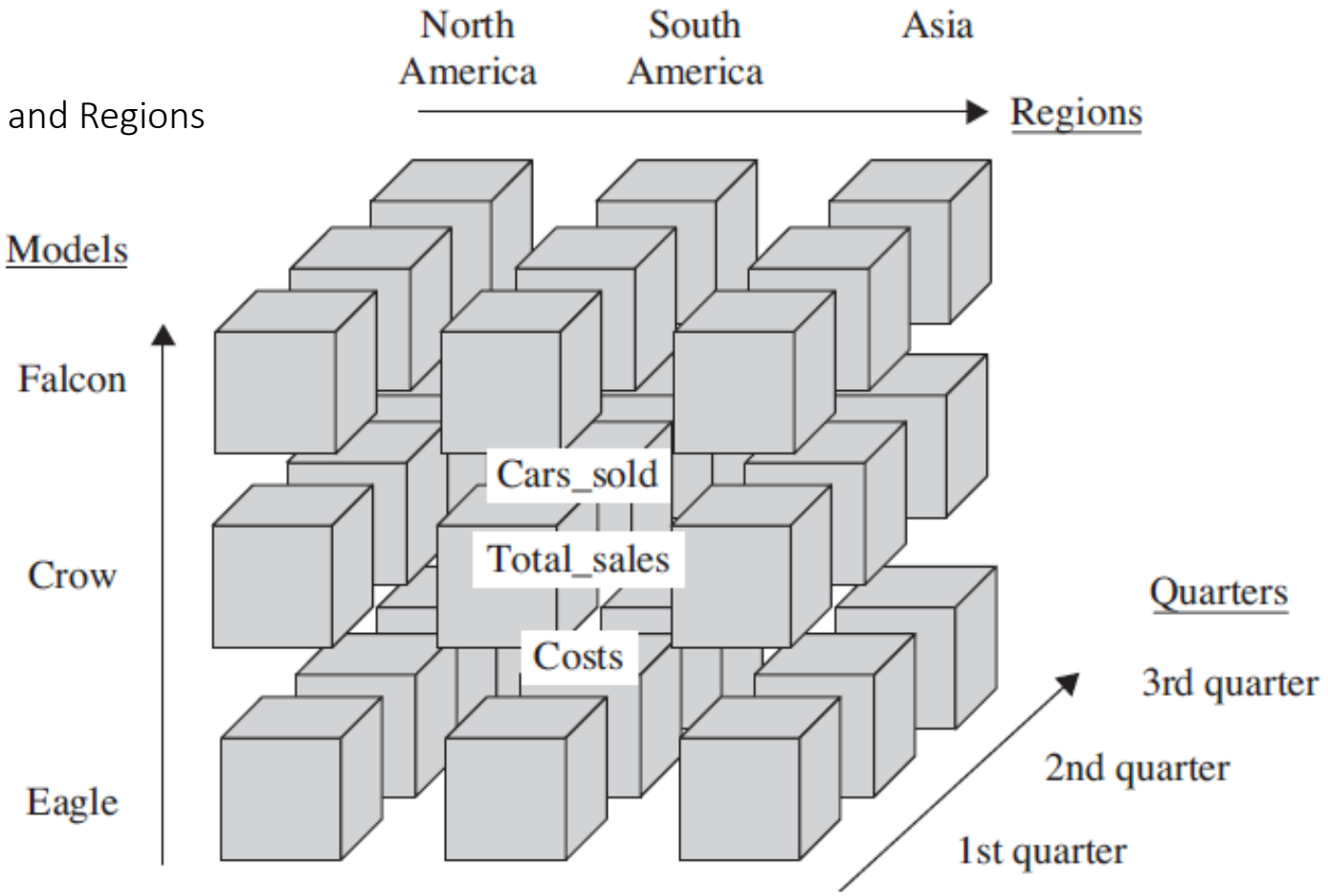
Cubes and their Architectures

RISTECH
RESEARCH AND INNOVATION
IN SCIENCE ENGINEERING AND TECHNOLOGY

- A *cube* is a subset of data from the data warehouse that can be organized into multidimensional structures.
- To define a cube, you first select a fact table from the dimensional schema and identify numerical columns (measures) of interest within it.
- Then you select dimension tables that provide descriptions for the set of data to be analyzed.
- Example (Car Sales Analysis)

Cube and their Architectures

Cube with dimensions Models, Quarters and Regions



Aggregation

- A typical query on a fact table fetches thousands or even millions of rows at a time, and the only useful operation upon such a huge amount of rows is to apply an aggregate function (sum, maximum, or average).
- Low-level data from the fact table should be summarized in advance and stored in intermediate tables.
- Such tables are called *aggregate tables*, and the whole process is called *aggregation*.

Physical Storage of Cube

- OLAP systems usually use one of the following three different architectures to store multidimensional data:
  - Relational OLAP (ROLAP)
  - Multidimensional OLAP (MOLAP)
  - Hybrid OLAP (HOLAP)

Data Access

- Data in a DWH can be accessed using one of the following three techniques:
  1. Reporting
  2. OLAP
  3. Data Mining
- Mining is the most complex of the three

DEMO

RISTECH
RESEARCH AND INNOVATION
IN SCIENCE ENGINEERING AND TECHNOLOGY

- SSAS (SQL Server Analysis Services) is a group of services that is used to manage data that is in a data warehouse or data mart.
- It organizes data from DWH into cubes with aggregates
- BIDS (Business Intelligence Development Studio) offer once interface for developing SSAS projects as well as SSIS and SSRS projects
- SSIS (SQL Server Integration Services)
- SSRS (SQL Server Reporting Services)

Demo (MS SSIS)

- OLTP Performance
  - The performance of a database system will increase if transactions in the database application programs are short.
  - The reason is that transactions use locks to prevent possible negative effects of concurrency issues.
- Large OLTP systems usually have many users working on the system simultaneously.
- A typical example is a reservation system for an airline company.

SSAS Terminology

- Cube
- Dimension (A *dimension* is a set of logically related attributes (stored together in a dimension table))
- Member (Each discrete value in a dimension is called a *member*)
- Hierarchy (specify groupings of multiple members within each dimension)
- Cell (*Cells* are parts of a multidimensional cube that are identified by coordinates)
- Level (describe the hierarchy from the highest level to the lowest level of data)
- Measure Group
- Partition