Date Warehousing

Lecturer: Neena Thota
neena.thota@it.uu.se

Where we are ……………….. Where we go

- **Database Design**
  - ER data modeling
  - Relational model
- **SQL**
  - DDL and DML
- **Database Control**
  - Transactions and ACID properties
  - Privileges with GRANTS
- **Database Tuning**
  - Normalization & data quality
  - Indexes and Queries

- **Data warehouses**
  - **Models**
  - **Operations**
  - **Architecture**

Data Mining

Data Analytics

BI

Intended learning outcomes

1. Understand need to **provide decision makers with information** at correct level of detail to support decision making;
2. Get familiar with **characteristics and functionalities** of data warehouses;
3. Understand how data is **modeled** in data warehouses;
4. Recognize **difficulties in implementing** data warehouses.

Introduction, Definitions, and Terminology

> **“A subject-oriented, integrated, nonvolatile, time-variant collection of data in support of management’s decisions.”**
>
> – W. H Inmon

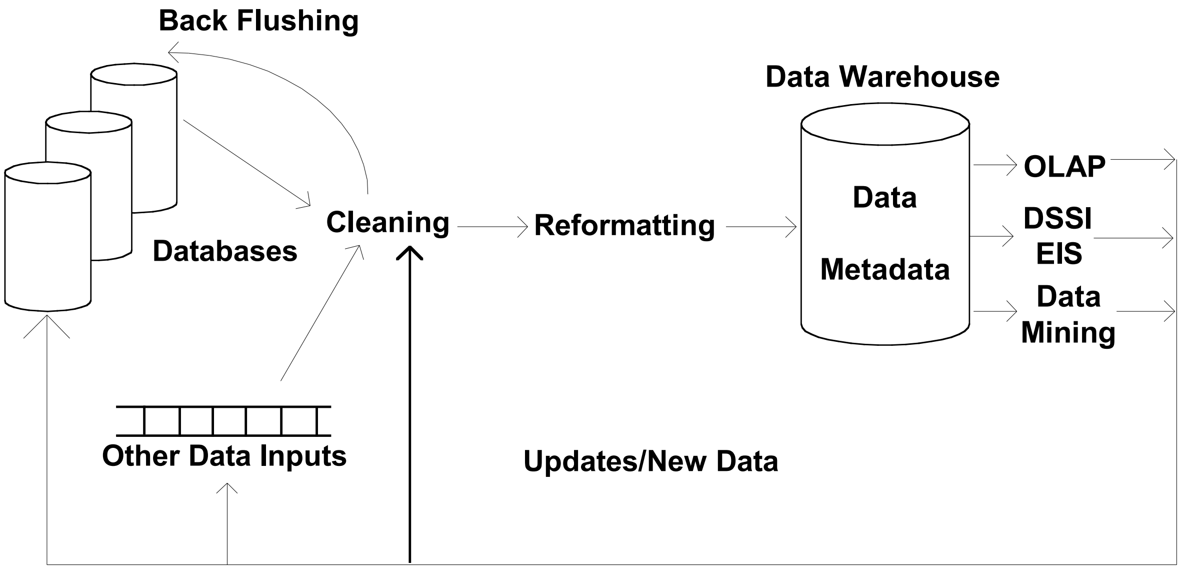
- **Subject oriented**:
  - targets one or several subjects of analysis according to analytical requirements of managers at various levels of the decision-making process.
- **Integrated:**
  - contents result from integration of data from various operational and external systems.
- **Nonvolatile:**
  - accumulates data from operational systems for a long period of time (data modification and removal not allowed; only operation allowed is purging of obsolete data no longer needed).
- **Time varying:**
  - keeps track of how its data has evolved over time.

Purpose of Data Warehousing

- Tools for decision makers to **make decisions quickly** and **reliably** based on historical data;
- Mainly intended for **decision support** applications;
- Users need only ***read access***
  - need the access to be ***fast*** over a ***large volume*** of data;
- Data comes from multiple databases
  - analysis are **recurrent and predictable** to be able to design specific software to meet the requirements.

Conceptual Structure of Data Warehouse

- **Cleaning and reformatting** of data
- **OLAP** (Online Analytical Processing): used to describe analysis of complex data.
- **DSS** (Decision Support Systems) also known as **EIS** (Executive Information Systems): supports organization's leading decision makers for making complex and important decisions.
- **Data Mining:** used for knowledge discovery- process of searching data for unanticipated new knowledge.



Comparison with Traditional Databases

Data Warehouse	Database
Mainly optimized for appropriate data access	Optimized for both access mechanisms and integrity assurance measures
Emphasize more on historical data as main purpose is to support time-series and trend analysis	Transaction based and volatile
Refresh policy is carefully chosen, usually incremental	Transaction is the mechanism change to the database

Classification of Data Warehouses

Enterprise-wide data warehouses

- Huge projects requiring massive investment of time and resources.

Virtual data warehouses

- Provide views of operational databases materialized for efficient access.

Data Marts

- Targeted to a subset of organization, such as a department, and are more tightly focused.

Data Modeling for Data Warehouses

- Traditional databases generally deal with two-dimensional data (similar to a spread sheet).
  - However, querying performance in a ***multi-dimensional data storage model*** is much more efficient.
- Data warehouses take advantage of this feature as generally these are
  - Non volatile;
  - The degree of predictability of the analysis that will be performed on them is high.

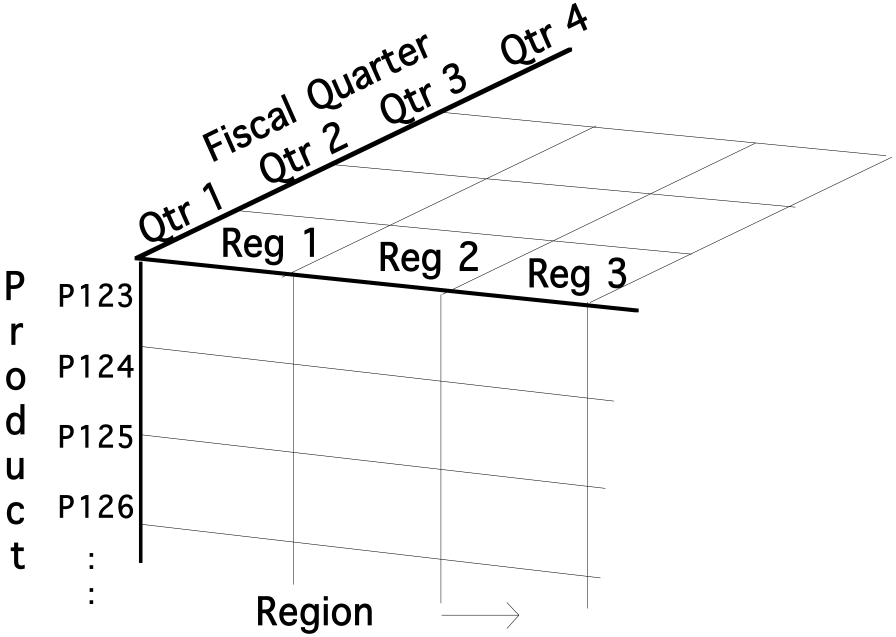
Two- Dimensional vs. Multi- Dimensional

Two Dimensional Model

REGION

PRODUCT		REG1	REG2	REG3
	P123			
	P124			
	P125			
	P126			
	:			
	:			

Three dimensional data cube

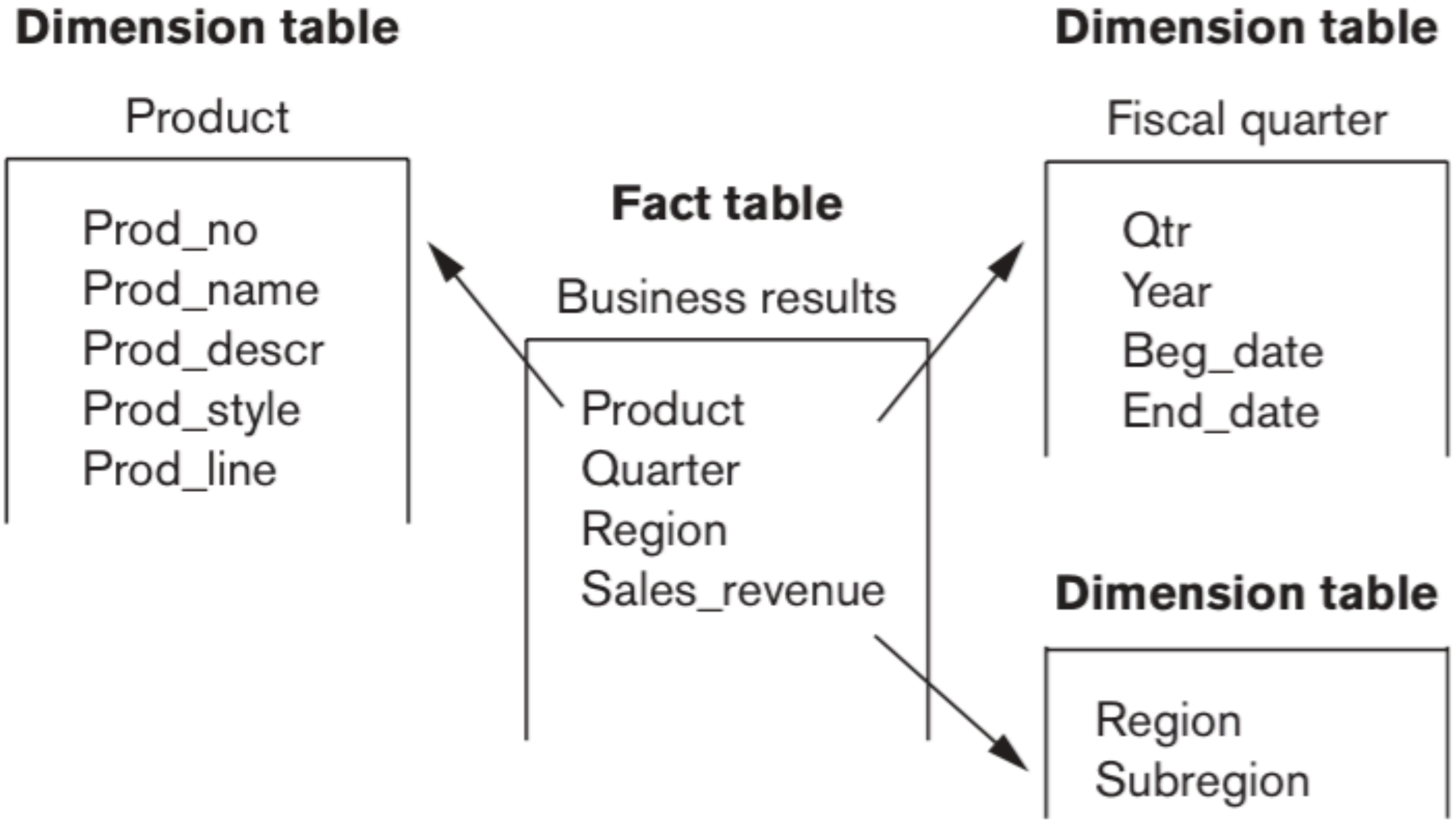


Multi-dimensional Schemas

- Specified using:
  - **Dimension table**
    - It consists of tuples of attributes of the dimension.
  - **Fact table**
    - Each tuple is a recorded fact. This fact contains some measured or observed variable (s) and identifies it with pointers to dimension tables. The fact table contains the data, and the dimensions to identify each tuple in the data.
  - Types:
    - **Star schema**
    - **Snowflake schema**

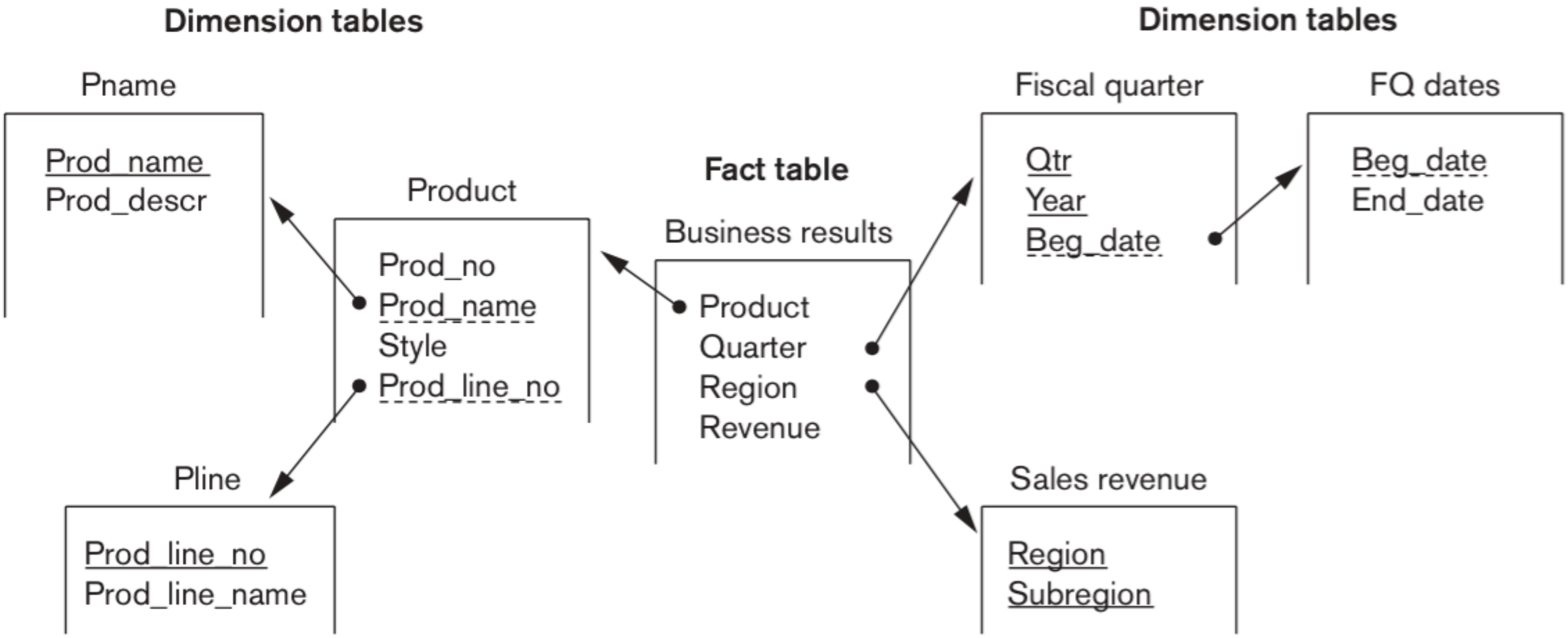
Star Schema

- Consists of a fact table with a single table for each dimension.



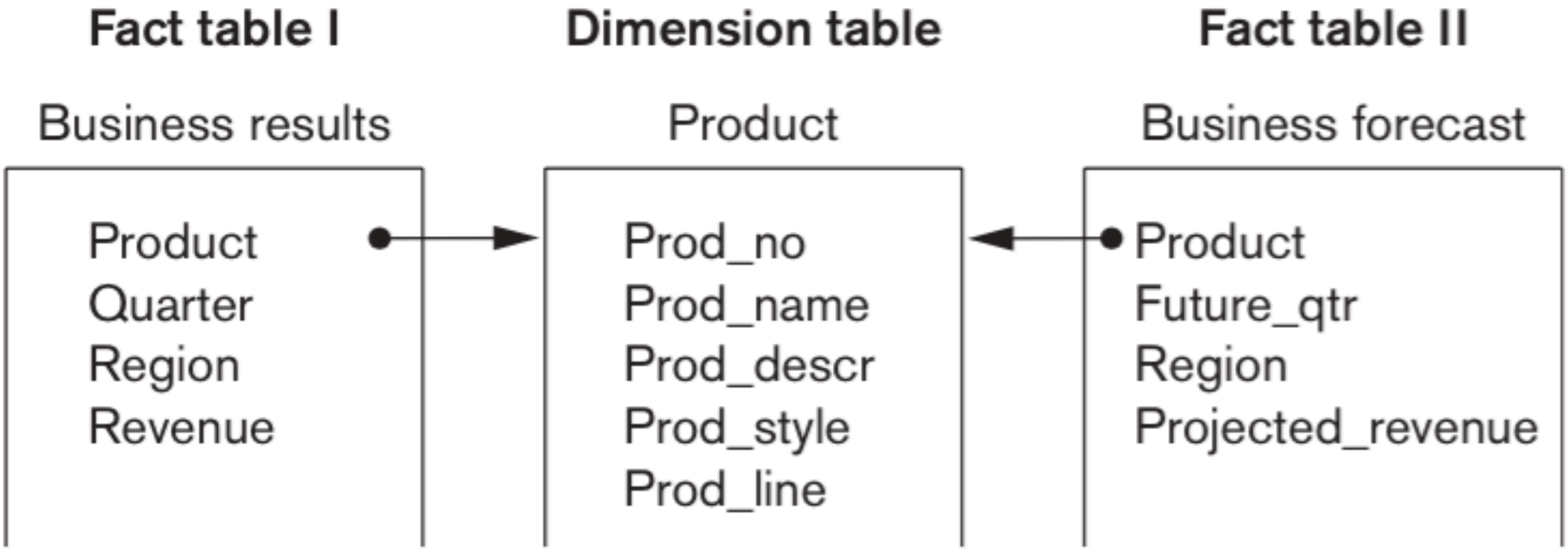
Snowflake Schema

- Variation of star schema
  - dimensional tables from star schema organized into a hierarchy by normalizing them.



Fact Constellation

- Set of tables that share some dimension tables.
  - Fact constellations limit possible queries for the warehouse.



Fact table I

Business results

Product
Quarter
Region
Revenue

Dimension table

Product

Prod_no
Prod_name
Prod_descr
Prod_style
Prod_line

Fact table II

Business forecast

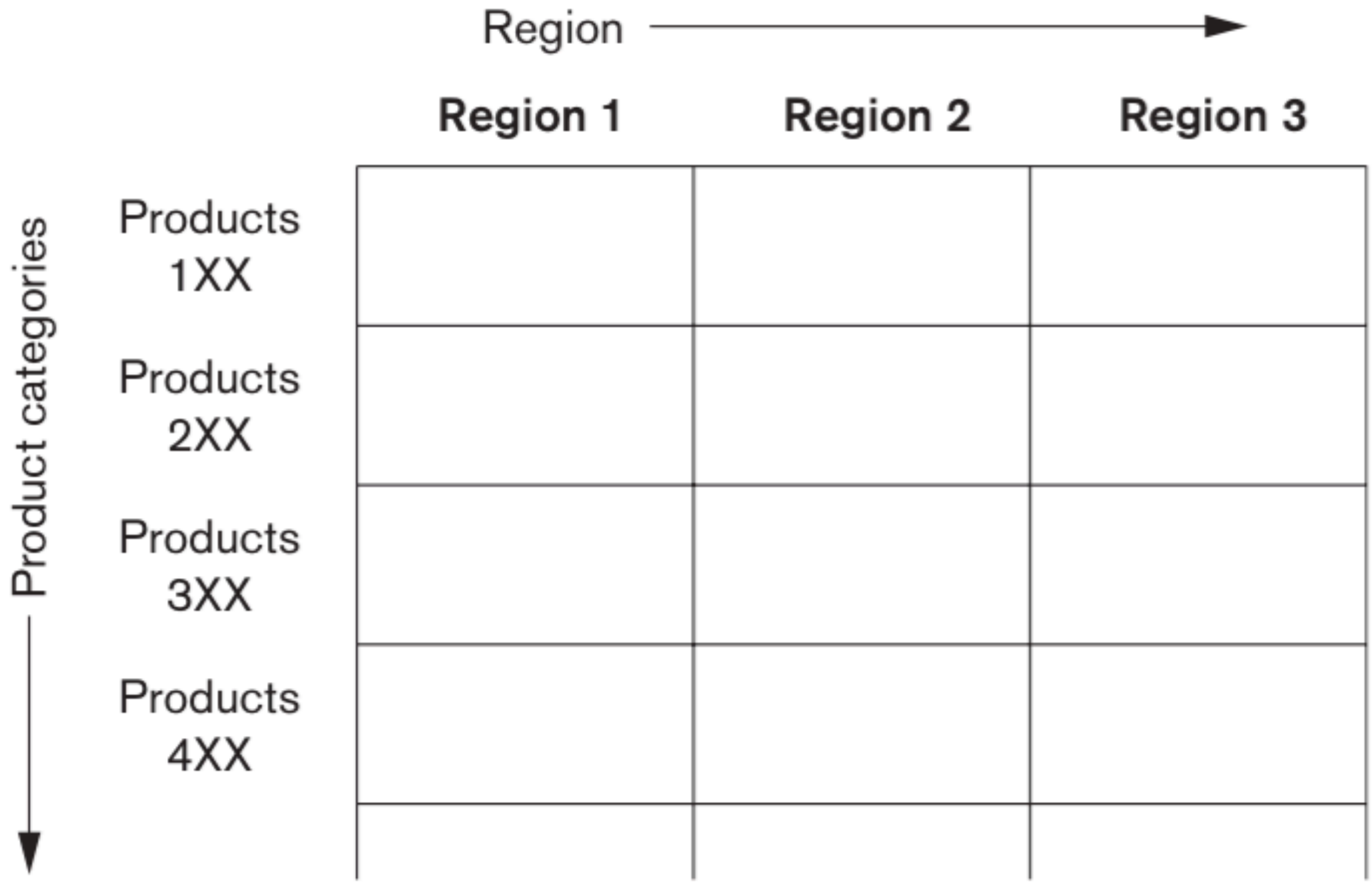
Product
Future_qtr
Region
Projected_revenue

Advantages of multi-dimensional models

- Lend themselves readily to **hierarchical views**:
    - Example: roll-up display, drill-down display;
    - Also allow cross-tabulation: pivot.
  - Data can be directly queried in any **combination of dimensions** bypassing complex database queries.

Roll-Up

- Data is summarized with increasing generalization



	Region 1	Region 2	Region 3
Products 1XX			
Products 2XX			
Products 3XX			
Products 4XX			

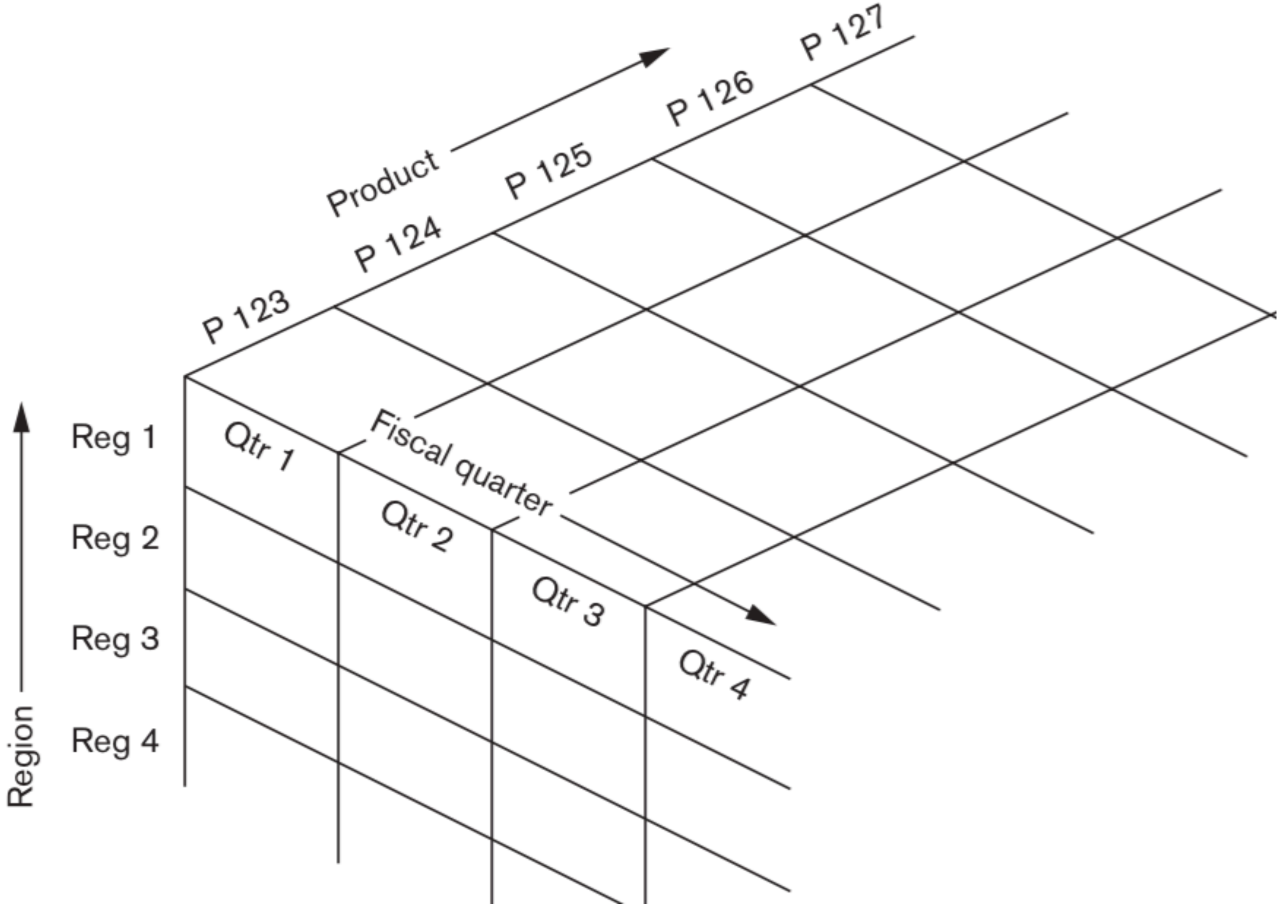
Drill-Down

- Increasing levels of detail are revealed

		Region 1				Region 2
		Sub_reg 1	Sub_reg 2	Sub_reg 3	Sub_reg 4	Sub_reg 1
P123 Styles	A B C D					
P124 Styles	A B C					
P125 Styles	A B C D					

Pivot

- Cross tabulation is performed.



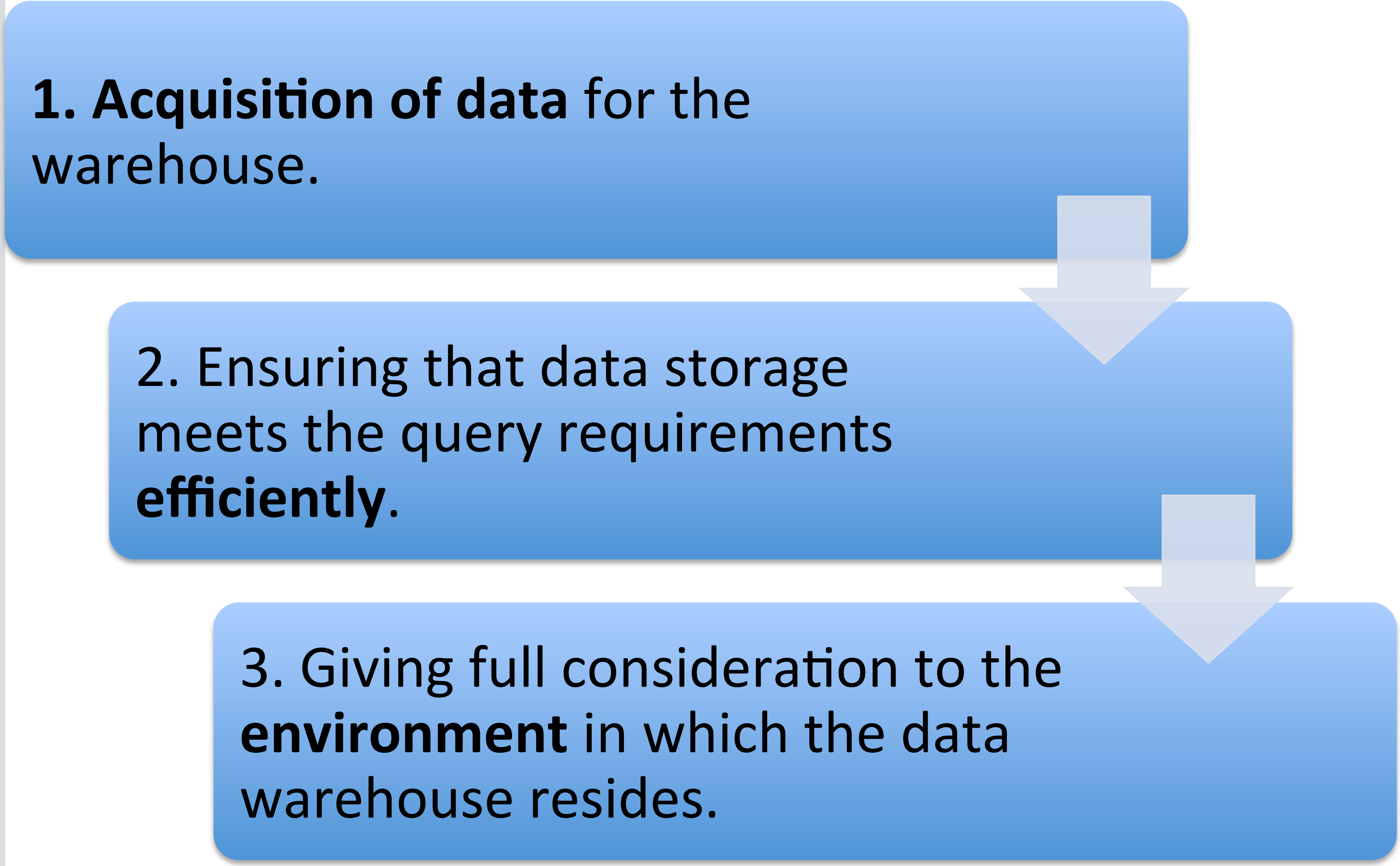
Other Functionality of a Data Warehouse

- **Slice and dice**:
  - Performing projection operations on the dimensions.
- **Sorting**:
  - Data is sorted by ordinal value.
- **Selection**:
  - Data is available by value or range.
- **Derived attributes**:
  - Attributes are computed by operations on stored derived values.

Indexing

- Data warehouse also utilizes indexing to support **high performance access**;
- A technique called **bitmap indexing** constructs a bit vector for each value in domain being indexed;
- Indexing works very well for domains of **low cardinality**.

Building A Data Warehouse – Design Steps



1. **Acquisition of data** for the warehouse.

2. Ensuring that data storage meets the query requirements **efficiently**.

3. Giving full consideration to the **environment** in which the data warehouse resides.

Step1: Acquisition of data

- Must be extracted from **multiple**, **heterogeneous** sources.
- Must be formatted for **consistency** within the warehouse.
- Must be **cleaned** to ensure validity.
  - Difficult to automate cleaning process.
  - Back flushing, upgrading the data with cleaned data.
- Must be fitted into **data model** of warehouse.
- Must be **loaded** into warehouse.
  - Proper design for refresh policy should be considered.

Step2: Ensuring data storage

- Storing the data according to **data model** of the warehouse
- Creating and maintaining required **data structures.**
- Creating and maintaining appropriate **access paths.**
- Providing for **time-variant data** as new data are added
- Supporting the **updating** of warehouse data.
- **Refreshing** the data
- **Purging** data

Step 3: Considerations of environment

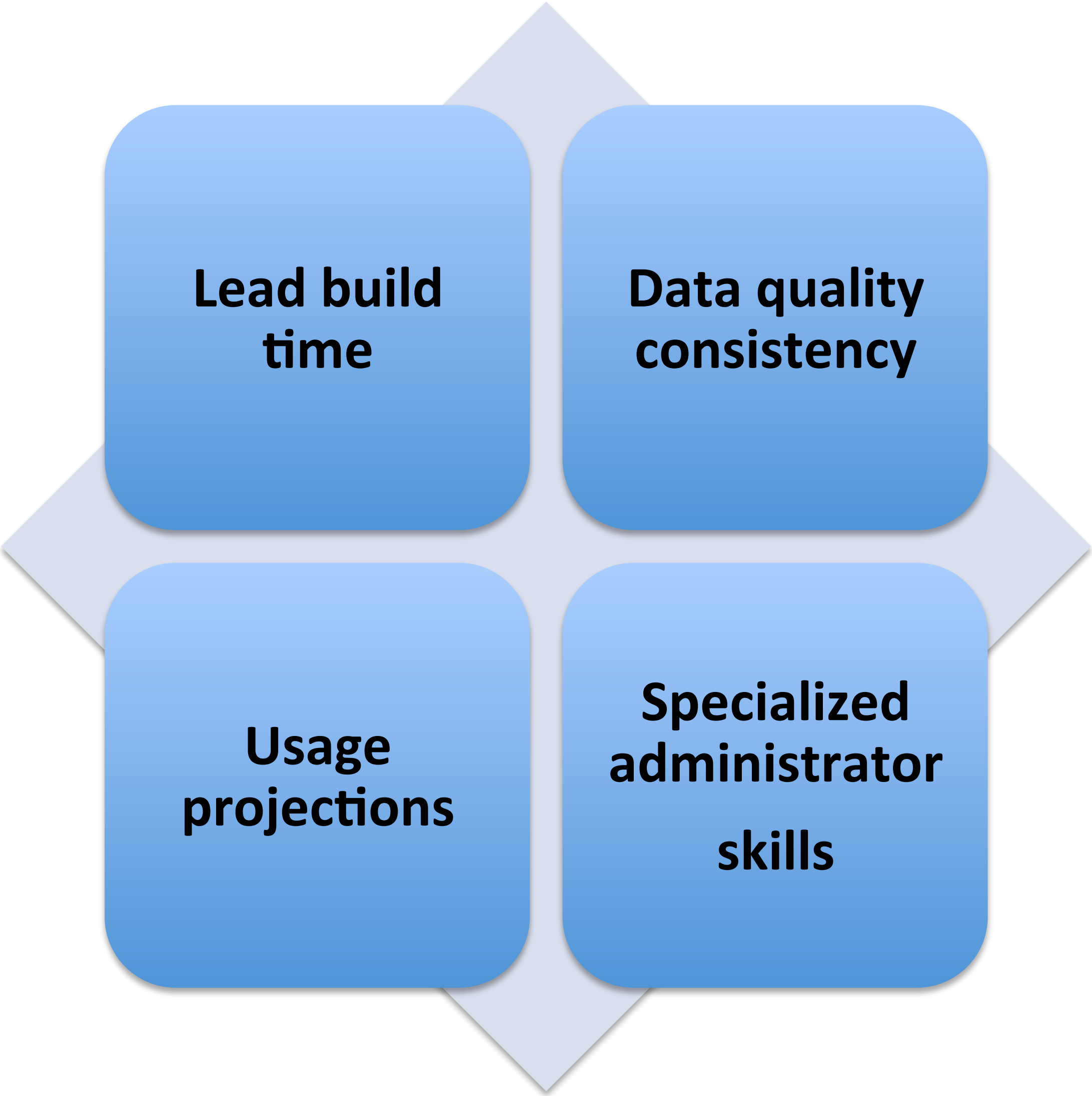
- **Usage** projections;
- The **fit** of the data model;
- Characteristics of available **resources**;
- Design of the **metadata** component;
- **Modular** component design;
- Design for **manageability and change;**
- Considerations of distributed and parallel architecture
  - Distributed vs. federated warehouses

Data Warehouse vs. Data Views

- Both have **read-only extracts** from the databases.

Data Warehouse	Views
Persistent storage	Materialized on demand
Multi-dimensional	Usually relational
Can index for optimization	Optimized
Specific support of functionality	No additional functionality
Huge volumes of data from many DBs	Usually one DB

Difficulties of implementing Data Warehouses



Lead build time

Data quality consistency

Usage projections

Specialized administrator skills

Issues in Data Warehousing

Data

- Cleaning
- Indexing
- Partitioning
- Views

Automation

- Data acquisition
- Quality management
- Access paths
- Functionality
- Performance optimization

Business Rules

- Domain rules

Summary

- Data warehouses supply decision makers with information at **correct level of detail**, based on an appropriate organization and perspective.
- Process that requires a variety of activities **to precede it**.
- Some special functionality associated with a multidimensional view of data:
  - Roll-up
  - Drill-down
  - Pivot…….
- Some difficulties and issues exist with usage.

Microsoft

TERADATA

ORACLE