UWE Bristol | University of the West of England

Presented by

Dr. Trupti Padiya

School of Computing and Creative Technologies

Advanced Databases UFCFU3-15-3

Temporal Databases, Data warehousing, Data marts, Data mining

Slides with modification adapted from Dr. Elias Pimenidis

Agenda

- Temporal Data and related concepts
- Data Warehousing
- Data Warehousing Operations
- Data Warehousing Applications
- Datamining

Temporal Data

- Temporal data is simply data that represents a state in time
- Timestamps
- Time intervals
- Time series data
- History of data
- Observation data, Sensor data, historical records, financial transaction
- A temporal database is a collection of time-referenced data. In such a database, the time references capture some temporal aspect of the data; put differently, the data are timestamped.

Data warehouse

- A subject-oriented, integrated, non-volatile, time-variant collection of data in support of management's decision making process.
- Data warehouses have the distinguishing characteristic that they are mainly intended for decision support applications.
- A data warehouse can be classed a temporal database. Most day-to-day databases are transactional databases.

A Temporal Data Warehouse

- A multidimensional data warehouse consists of three different levels:
- The schema level (dimensions, categories), the instance level (dimension members, master data), and the data level (data cells, transaction data).
- *Data Warehouse Maintenance:* The (ongoing) process and methodology of performing changes on the schema and instance level to represent changes in the data warehouse's application domain or requirements.
- *Data Warehouse Evolution* is a form of data warehouse maintenance where only the newest data warehouse state is available.
- *Data Warehouse Versioning* is a form of data warehouse maintenance where all past versions of the data warehouse are kept available.
- Dealing with changes on the data level, mostly insertion of new data, is not part of data warehouse maintenance, but part of a data warehouse's normal operation.

Purpose of Data Warehousing

- Most of the times the data warehouse users need **only read access** but, need the access to be fast over a large volume of data.
- Most of the data required for **data warehouse analysis** comes from multiple databases and such analysis is recurrent and predictable so that specific software can be designed to meet the requirements.
- There is a great need for tools that provide decision makers with information to make decisions quickly and reliably based on **historical data**.

Data Warehouses are used for

- **OLAP** (Online Analytical Processing) is a term used to describe the analysis of complex data from the data warehouse.
- **DSS** (Decision Support Systems) also known as EIS (Executive Information Systems) supports leading decision makers of organizations for making complex and important decisions.
- **Data Mining** is used for knowledge discovery, the process of mining data for unanticipated new knowledge.

Comparison with Traditional Databases

- Data Warehouses are mainly **optimized** for appropriate data access.
  - Traditional databases are transactional and are optimized for both access mechanisms and integrity assurance measures.
- Data warehouses **emphasize more on historical data**
  - main purpose is to support time-series and trend analysis
- Compared with transactional databases, data warehouses are **nonvolatile**.
  - In transactional databases **transactions** are the mechanism change to the database.
  - By contrast information in data warehouse is relatively **coarse grained** and refresh policy is carefully chosen, usually incremental.

Critically think on other perspectives about comparing data warehouse with traditional databases.

Classification of Data warehouse

- Generally, Data Warehouses **are an order of magnitude larger** than the source databases.
- The sheer volume of data is an issue, based on which Data Warehouses could be **classified** as follows.
  - **Enterprise-wide data warehouses**
    - They are huge projects requiring massive investment of time and resources.
  - **Virtual data warehouses**
    - They provide views of operational databases that are materialized for efficient access.
  - **Data marts**
    - These are generally targeted to a subset of organization, such as a department, and are more tightly focused.

Decision Support System

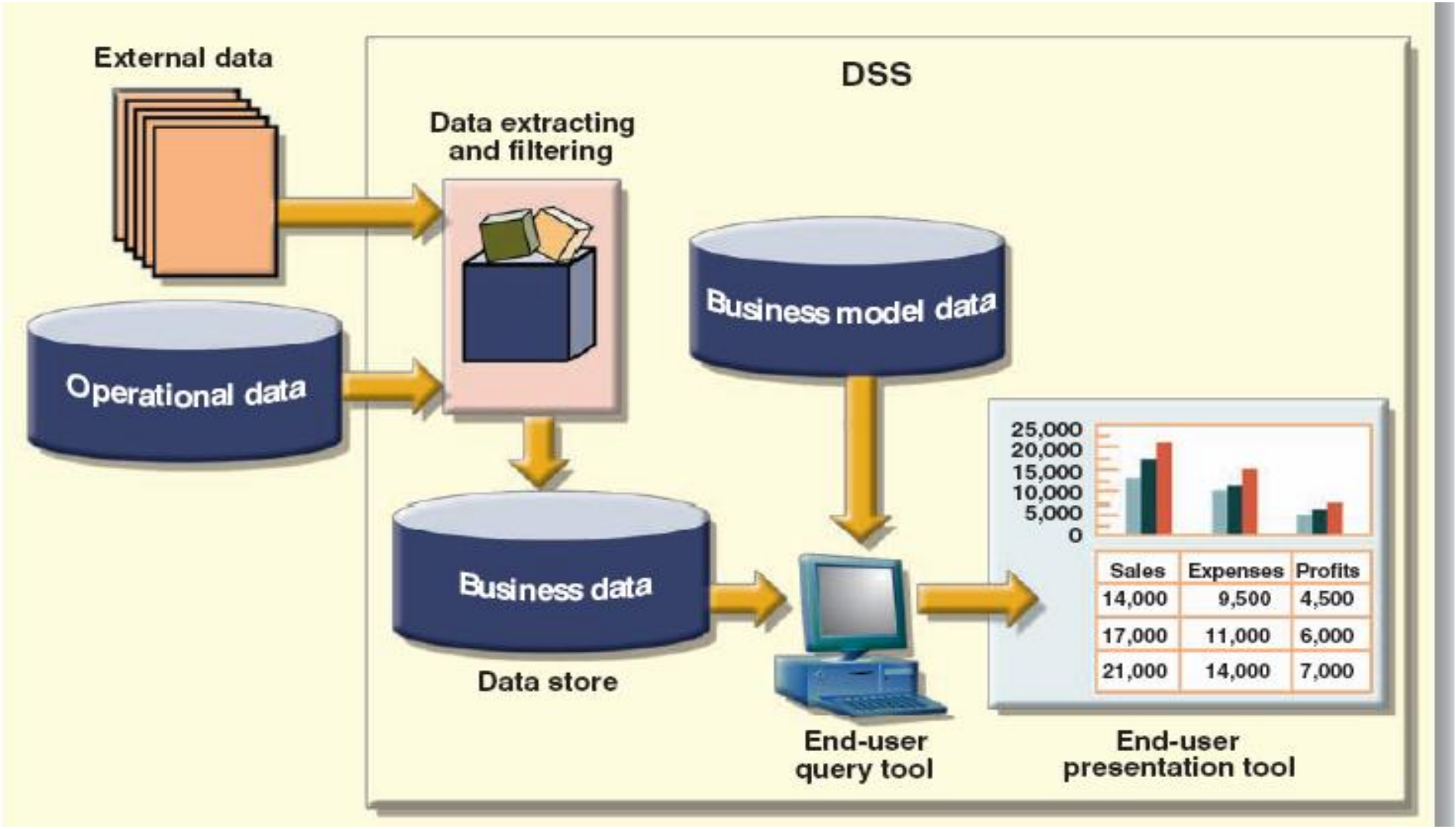
- Decision support is a methodology (or a series of methodologies) designed to extract information from data and to use such information as a basis for decision making
- Decision support system (DSS)
  - Arrangement of computerized tools used to assist managerial decision making within business
  - Usually requires extensive data “massaging” to produce information
  - Used at all levels within organization
  - Often tailored to focus on specific business areas
  - Provides query tools to retrieve data and to display data in different formats

Decision Support Systems Components

DSS are composed of following four main components:

- Data store component - Basically a DSS database
- Data extraction and data filtering component. Used to extract and validate data taken from operational database and external data sources
- End-user query tool. Used to create queries that access database
- End-user presentation tool. Used to organize and present data

Decision Support System – Main Components



Operational data vs Decision support data

- Operational Data
  - Mostly stored in relational database
  - Optimized to support transactions representing daily operations
- DSS Data
  - Give **tactical and strategic** business meaning to operational data
  - Differs from operational data in following three main areas:
    - Timespan – day to day VS Summarised
    - Granularity – level of detail (e.g. how a name field is subdivided)
    - Dimensionality – number of variables related to the data

Data Mart

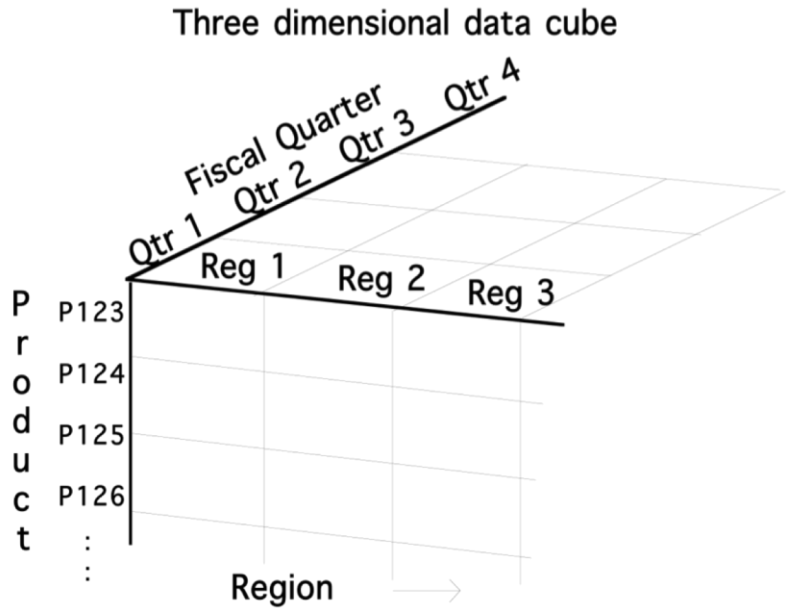
- Small, single-subject data warehouse subset
- Each has a more manageable data set than a data warehouse
- Provides decision support to small group of people
- Typically lower cost and lower implementation time than a data warehouse

Data Modelling for Data warehouse

- 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.

Two Dimensional Model

PRODUCT	REGION: REG1	REG2	REG3
P123			
P124			
P125			
P126			
⋮			



Multi-dimensional Models

- A data model is composed of a fact table with a *composite primary key* and a set of dimension tables
  - **Dimensions** – Product, Time, Region
  - **Facts** – Data represented by the cube e.g. sales across certain dimensions
- **Fact Table**
  - Stores business facts. 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
- **Dimension Table**
  - Contains information on one dimension
  - Consists of tuples of attributes of the dimension

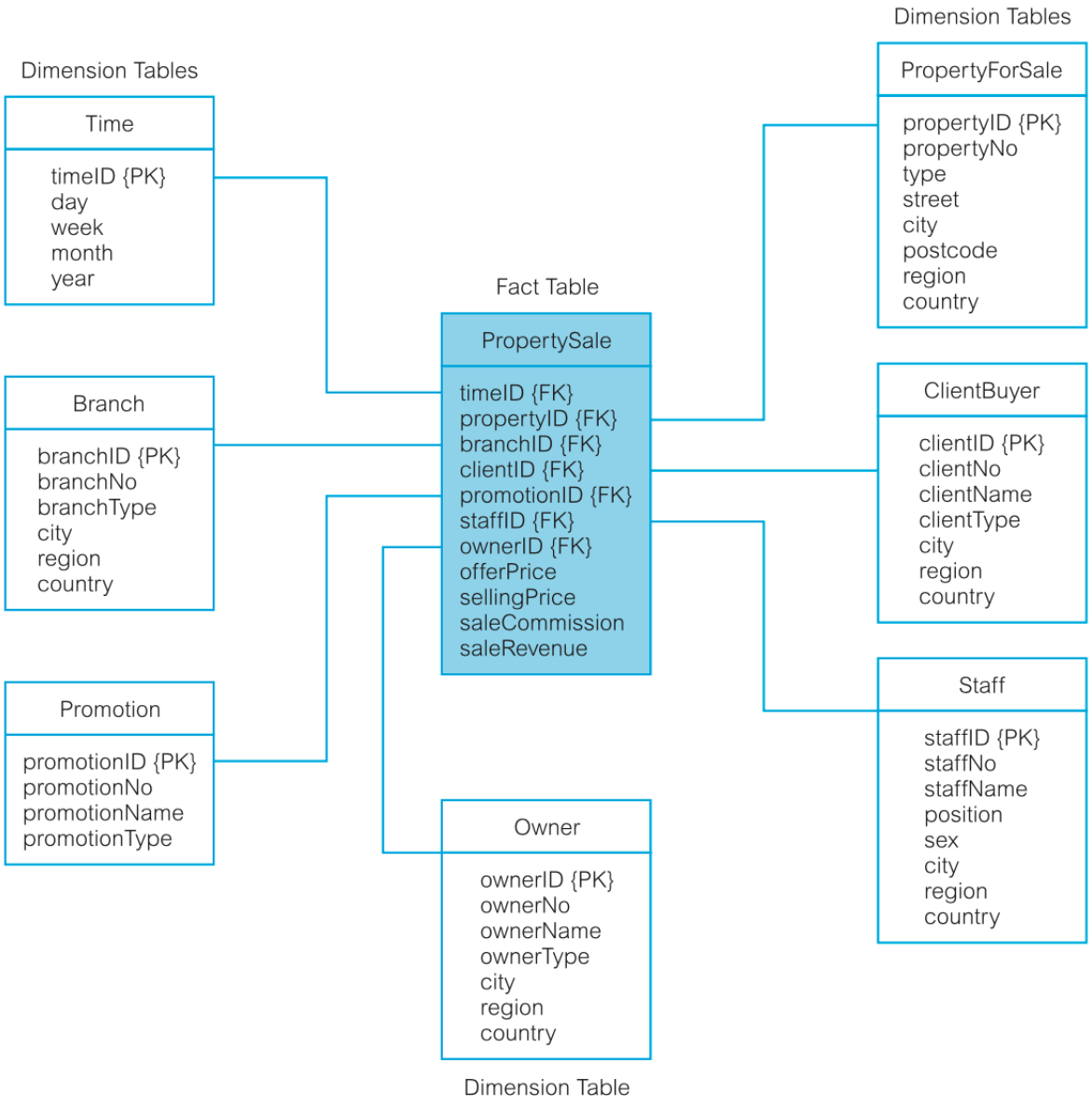
Multi-dimensional Schema

- Two common multi-dimensional schemas are
  - **Star Schema**:
    - A multi-dimensional data model that consists of a fact table in the centre, surrounded by *de-normalised* dimension tables
  - **Snowflake Schema**:
    - Consists of a fact table in the centre, surrounded by *normalised* dimensional tables
    - A variation of star schema, in which the de-normalised dimension tables from a star schema are organized into a hierarchy by normalising them.

Star Schema

- Star schemas can be used to speed up query performance by de-normalising reference data into a single dimension table
- e.g. in the star schema for property sales on next slide, PropertyForSale, Branch, ClientBuyer, Staff, and Owner all contain location data (city, region, and country), which is repeated in each

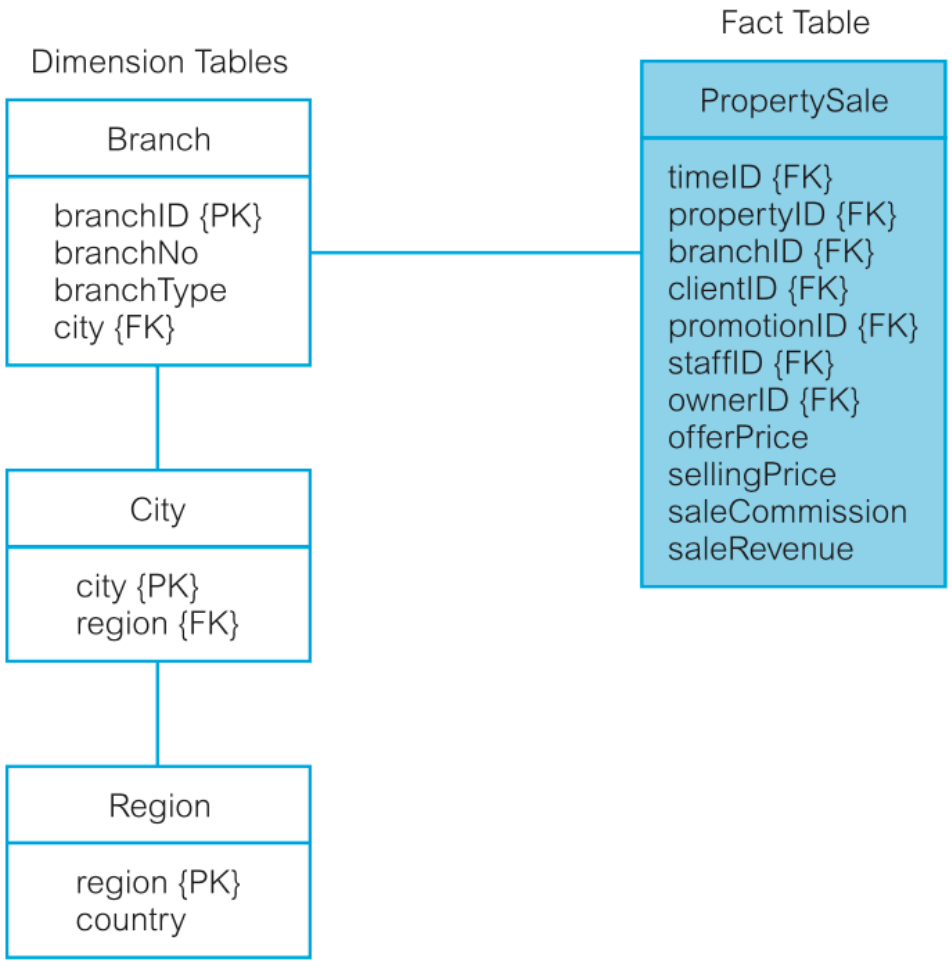
Star Schema for Property Sales



Snowflake Schema

- It is a variation of star schema, in which the dimensional tables from a star schema are organized into a hierarchy by normalizing them.
- e.g. we could normalise the location data (city, region, and country attributes) in the Branch dimension table in the star schema for property sales to create two new dimension tables called City and Region
- Normalisation is appropriate where the additional data is not accessed very often, the overhead of scanning the expanded dimension table may not be offset by any gain in the query performance

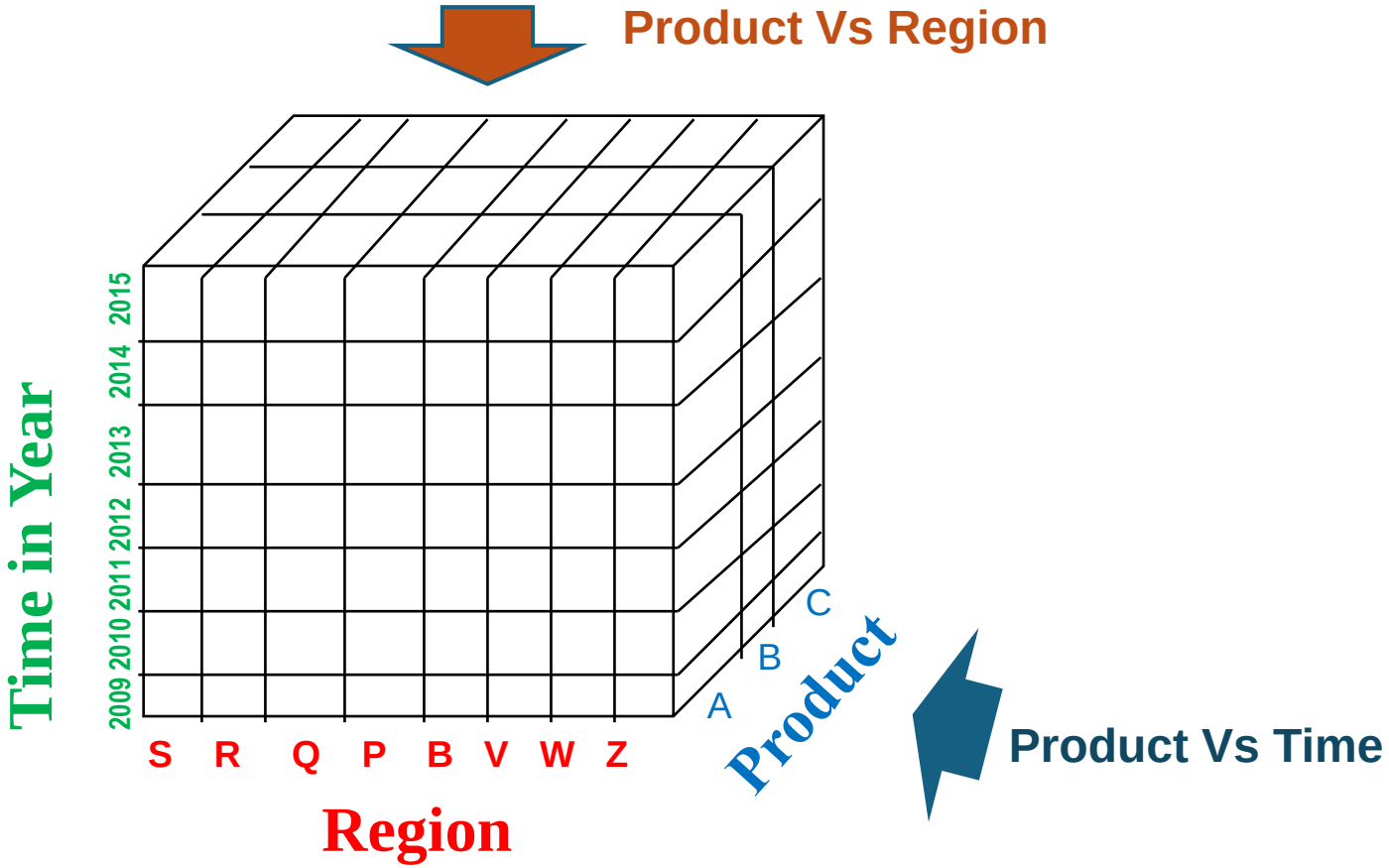
Snowflake Schema



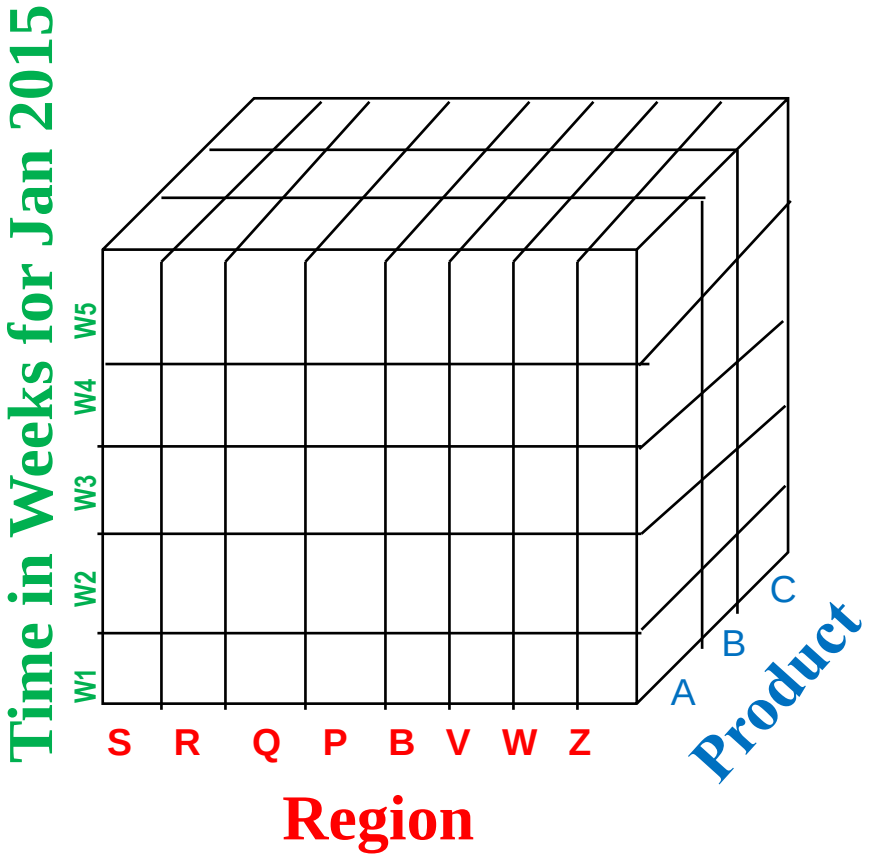
Operations of a Data Warehouse

- **Roll-up**: Data is aggregated with increasing generalization
  - Moving up the dimensional hierarchy (e.g. postal code to area to city)
  - Dimension reduction (e.g. from location, time, type, office to location, time, type)
- **Drill-Down**: Reverse of roll-up involving revealing the detailed data that forms the aggregated data
- **Slice and dice**: Performing projection operations on the dimensions
  - Slice – selecting one dimension
  - Dice – selecting two or more dimensions
- **Pivot**: Rotating the data/dimensions

Operations in a multi-dimensional model - Pivot



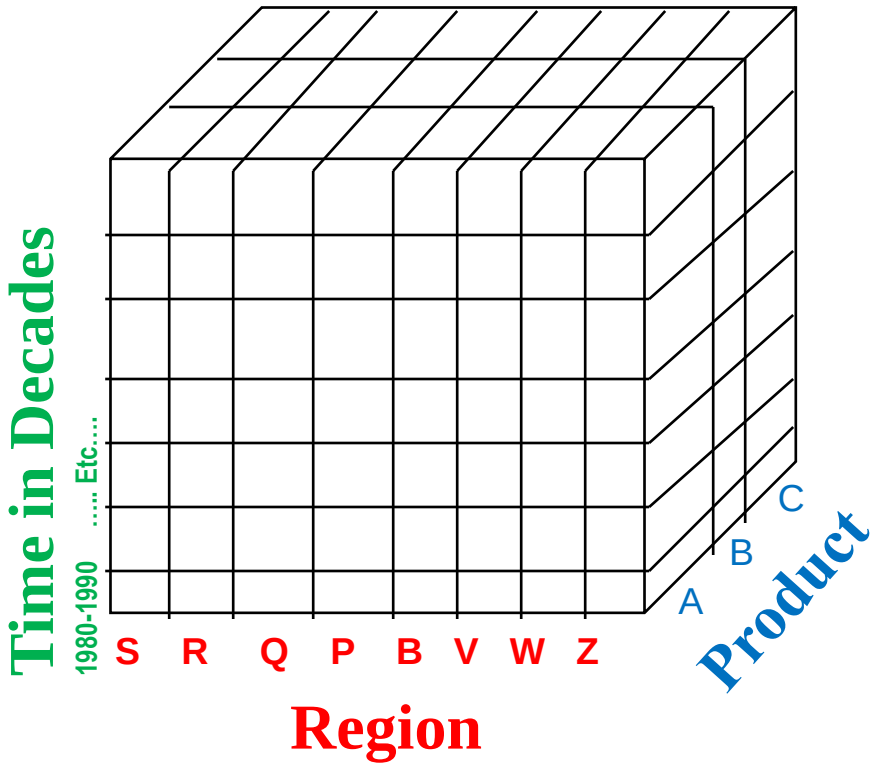
Operations in multi-dimensional model – Drill down



Operations in a multi-dimensional model – Roll up

- **Roll-up**

  e.g. Time roll-up to decade, Product rollup to brand



Data Mining

- Data mining is the process of uncovering hidden and potentially useful information from data.
- Two main types of data mining
  - Descriptive data mining
    - Discover patterns in data that are human interpreted.
  - Predictive data mining
    - Predict the value (a class or numerical) of an attribute using values of other attributes.

Data Mining Tasks

- Descriptive data mining
  - Clustering
  - Association Rule Discovery
  - Sequential Pattern Discovery
- Predictive data mining
  - Classification
  - Regression
  - Deviation Discovery

Data Mining Tasks

- **Classification** – Group Records in a Class
  - Example: classify bank customers based on credit rating (risky, average, excellent).
  - **Input** to the algorithm = a set of data describing objects that have already been classified
    - known as training data or training examples
  - **Output** = a model that can be used to classify other objects
    - different algorithms produce different types of models

Example: Medical Diagnosis

- Given a set of symptoms, we want to be able to determine a correct diagnosis for a patient with cold-like symptoms.
- Sample training data

Patient ID#	Sore Throat	Fever	Swollen Glands	Congestion	Headache	Diagnosis
1	Yes	Yes	Yes	Yes	Yes	Strep throat
2	No	No	No	Yes	Yes	Allergy
3	Yes	Yes	No	Yes	No	Cold
4	Yes	No	Yes	No	No	Strep throat
5	No	Yes	No	Yes	No	Cold
6	No	No	No	Yes	No	Allergy
7	No	No	Yes	No	No	Strep throat
8	Yes	No	No	Yes	Yes	Allergy
9	No	Yes	No	Yes	Yes	Cold
10	Yes	Yes	No	Yes	Yes	Cold

Rule for Classification

- One possible model that could be used for classifying other patients is a set of rules like the following:

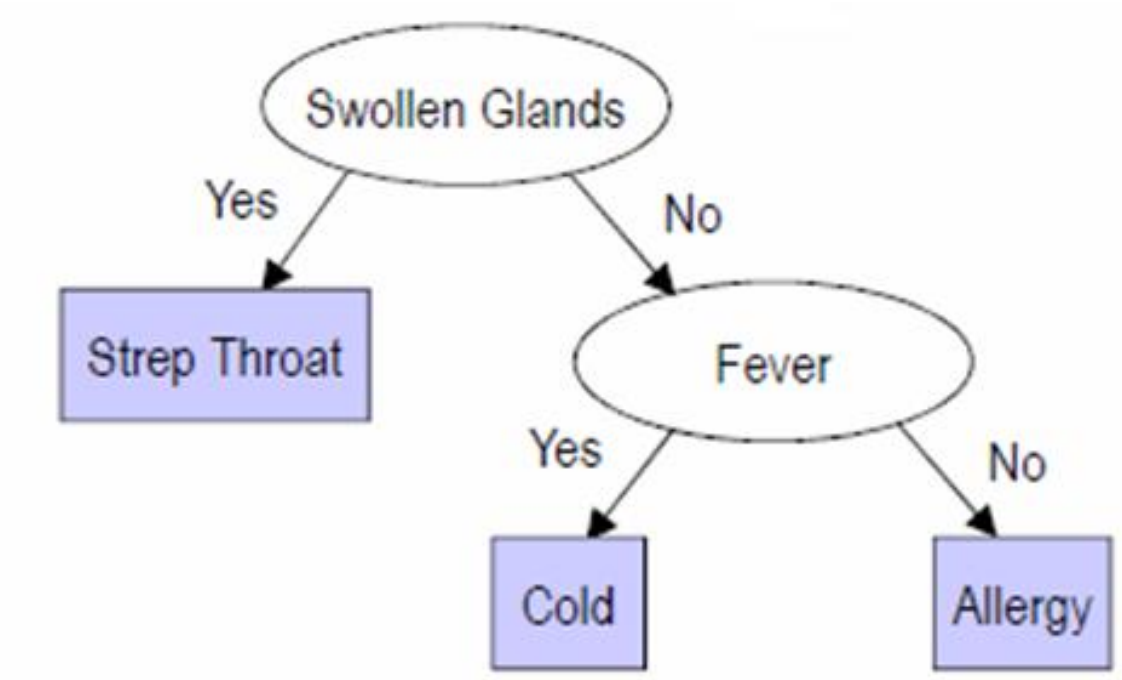
if Swollen Glands == Yes
then Diagnosis = Strep Throat

if Swollen Glands == No and Fever == Yes
then Diagnosis = Cold

if Swollen Glands == No and Fever == No
then Diagnosis = Allergy

Decision Trees for classification

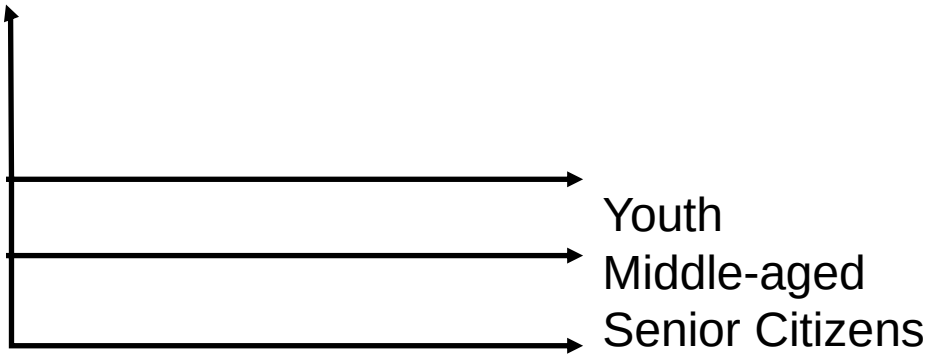
- Another possible type of model is known as a decision tree:



Data Mining Tasks

- **Clustering** – Try to group records using some similarity measures.
  - Example, we can classify *Customers* into different categories using the following measures

- Time spent on internet
- Time texting
- Café Time
- Movies



Can use this information if I want to target a particular group of people for a particular product.

Data Mining Tasks

- **Association Rule Discovery**
  - Trying to predict which set of items go together or can happen together.
- Example: in a supermarket, try to predict which items customers buy together.

{Milk, Diapers} → {infant food}

{Milk, Bread} → {eggs}

Data Mining Tasks

- **Sequential Pattern Discovery**
  - Finding out dependencies between events in time
- Example, in a sports shop, the following is a pattern of purchases.

  (Bat) (Shoes) → (Leg Pads)

Data Mining Tasks

- **Deviation**
  - Finding abnormal behaviour
- Example

  - Person A spends £400 on credit card each week.

  - Then Person A buys £20,000 worth of stuff every 2 days.

Significant Deviation in behaviour

Data Mining Challenges

- Noisy and Incomplete Data
- Distributed Data
- Complex Data
- Bias
- Heterogeneity
- Incorporation of Background Knowledge
- Data Privacy and Security

Some extra resources

- https://www.ibm.com/think/topics/data-warehouse
- https://learn.microsoft.com/en-us/power-bi/transform-model/datamarts/datamarts-overview
- https://www.ibm.com/think/topics/data-mining