MSc Computing- Enterprise Software Systems

Organisation Memory II

Lecturer: Ruth Barry
Department: Computing and Mathematics
Email: rbarry@wit.ie
Website: www.wit.ie

Waterford Institute *of* Technology
INSTITIÚID TEICNEOLAÍOCHTA PHORT LÁIRGE

Objectives

Creating a Data Warehouse

Why separate a Data Warehouse?

Three kinds of DW applications

The Role of a Business Analyst

Multidimensional Data Representation and Manipulation

Data Warehouse design Implementation – things to avoid…

Important applications and trends in DW technology

Traditional Applications - examples

Industry	Key Applications
Airline	Yield management, route assessment
Telecommunications	Customer retention, network design
Insurance	Risk assessment, product design, fraud detection
Retail	Target marketing, supply-chain management

Tools/Applications that Connect with DW:

Information processing –

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

Analytical processing –

- multidimensional analysis of data warehouse
- 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.

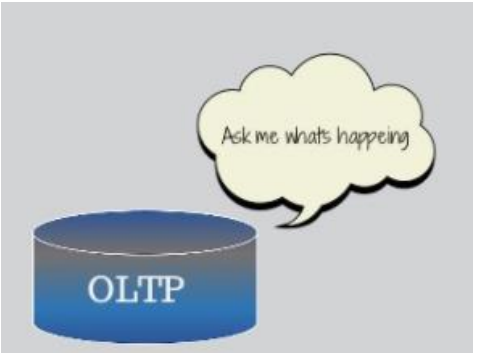
Creating a Data Warehouse

What do you want to know?

Data Warehouse vs Operational Database

Operational database – e.g. transaction processing system. Also known as Online Transaction Processing System OLTP

- Major task of traditional relational DBMS – Day-to-day operations: purchasing, inventory, banking, manufacturing, payroll, registration, accounting, etc.



Data Warehouse – is to provide analysis. Online Analytical Processing (OLAP)

- OLAP connects to a data warehouse system –
  - Data analysis and decision making
  - Can organize and present data in various forms and combinations

Why separate a 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

OLAP vs. OLTP

Criteria	OLTP	OLAP
Purpose	To carry out day-to-day business functions	To support decision making and provide answers to business and management queries
Data source	Transaction database (a normalized data repository primarily focused on efficiency and consistency)	Data warehouse or DM (a nonnormalized data repository primarily focused on accuracy and completeness)
Reporting	Routine, periodic, narrowly focused Reports	Ad hoc, multidimensional, broadly focused reports and queries
Resource requirements	Ordinary relational databases	Multiprocessor, large-capacity, specialized databases
Execution speed	Fast (recording of business transactions and routine reports)	Slow (resource intensive, complex, large-scale queries)

OLAP Operations

Slice - a subset of a multidimensional array

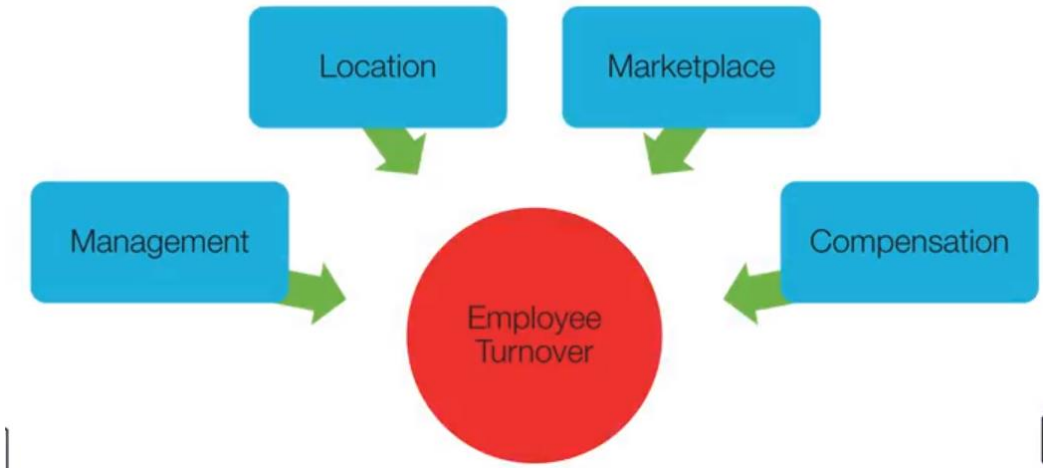
Dice - a slice on more than two dimensions

Drill Down/Up - navigating among levels of data ranging from the most summarized (up) to the most detailed (down)

Roll Up - computing all of the data relationships for one or more dimensions

Pivot - used to change the dimensional orientation of a report or an ad hoc query-page display

The Role of a Business Analyst



People who work with accessing data in a data warehouse are known as Business analysts.

Business analysts typically think about problems from a perspective of factors – such as location, impacting an outcome variable, such as an employee turnover

Early developers of DW software developed a data model, known as, “data cube”, to support this type of reasoning.

Data Cube Basics

Business analyst

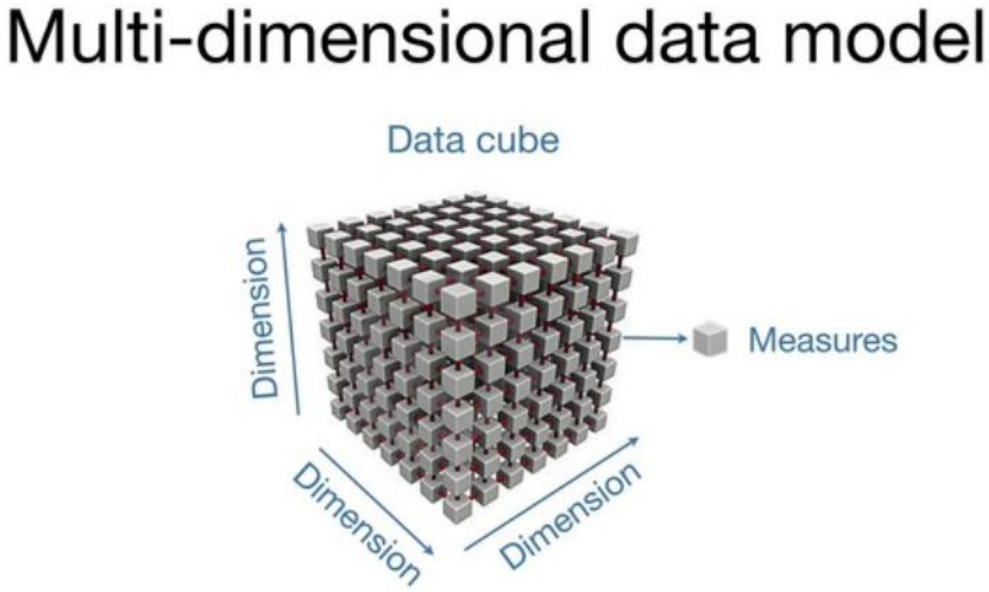
- Starts with factors or influencing variables of interest
- Quantitative variables – e.g. unit
- Multidimensional arrangement

Steps involved are:

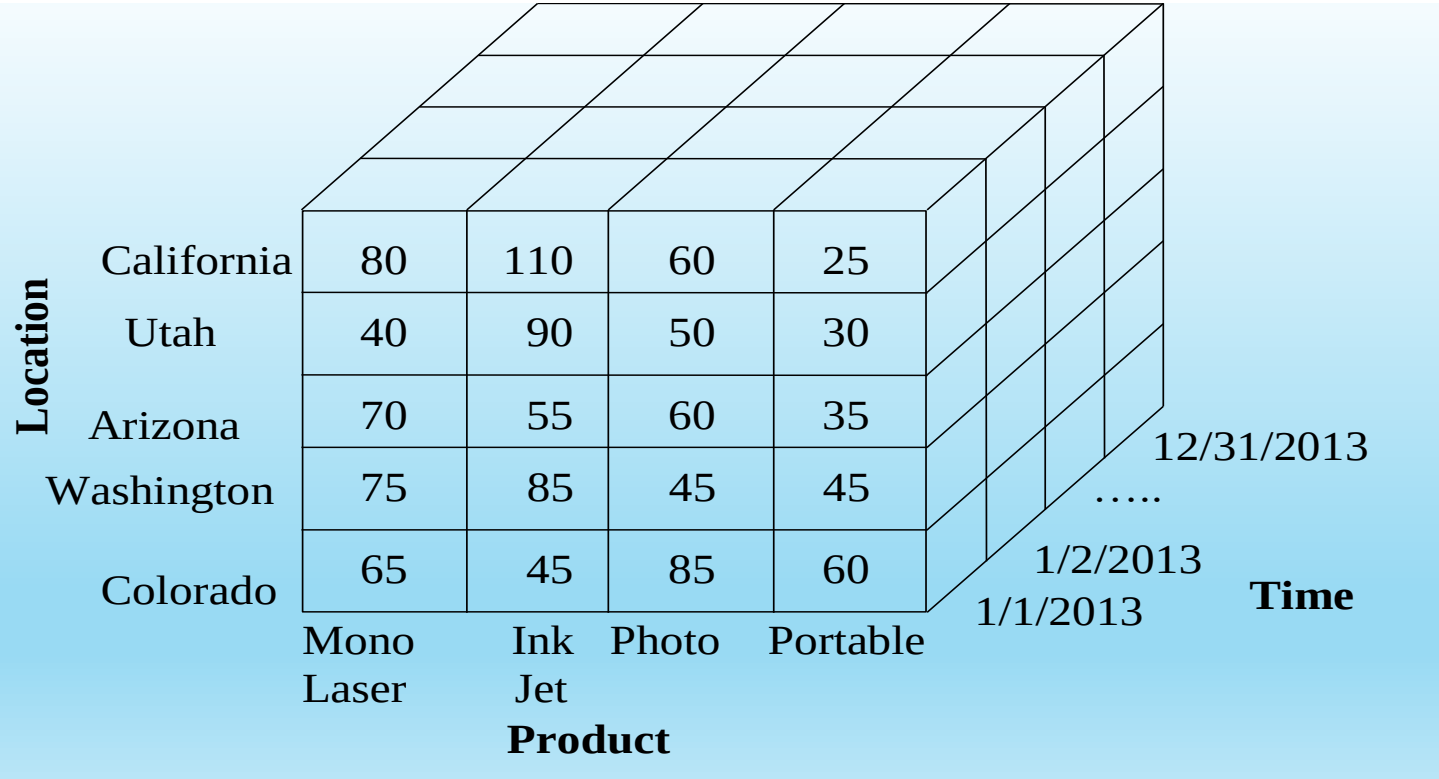
- Source – create data view – create cube

Terminology

- Dimension: subject label for a row or column
  - Dimensions are organized into hierarchies
  - Dimensions have attributes
- Measure: quantitative variables stored in cells

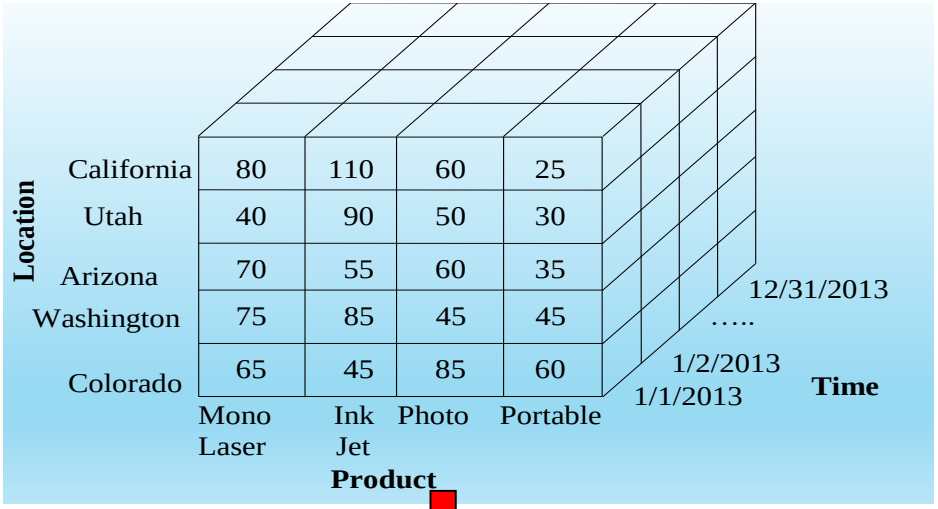


Sales Data Cube Example



Slice Operator

- Subset of dimensions
- Set dimension to specific value



(Location × Product Slice for Time = 1/1/2013)

Location	Product			
	Mono Laser	*Ink Jet*	*Photo*	*Portable*
California	80	110	60	25
Utah	40	90	50	30
Arizona	70	55	60	35
Washington	75	85	45	45
Colorado	65	45	85	60

Dice Operator

- Replace a dimension with a subset of values
- Dice operation often follows a slice operation

Location	Product			
	Mono Laser	*Ink Jet*	*Photo*	*Portable*
California	80	110	60	25
Utah	40	90	50	30
Arizona	70	55	60	35
Washington	75	85	45	45
Colorado	65	45	85	60

(Utah, Colorado, Arizona Dice)

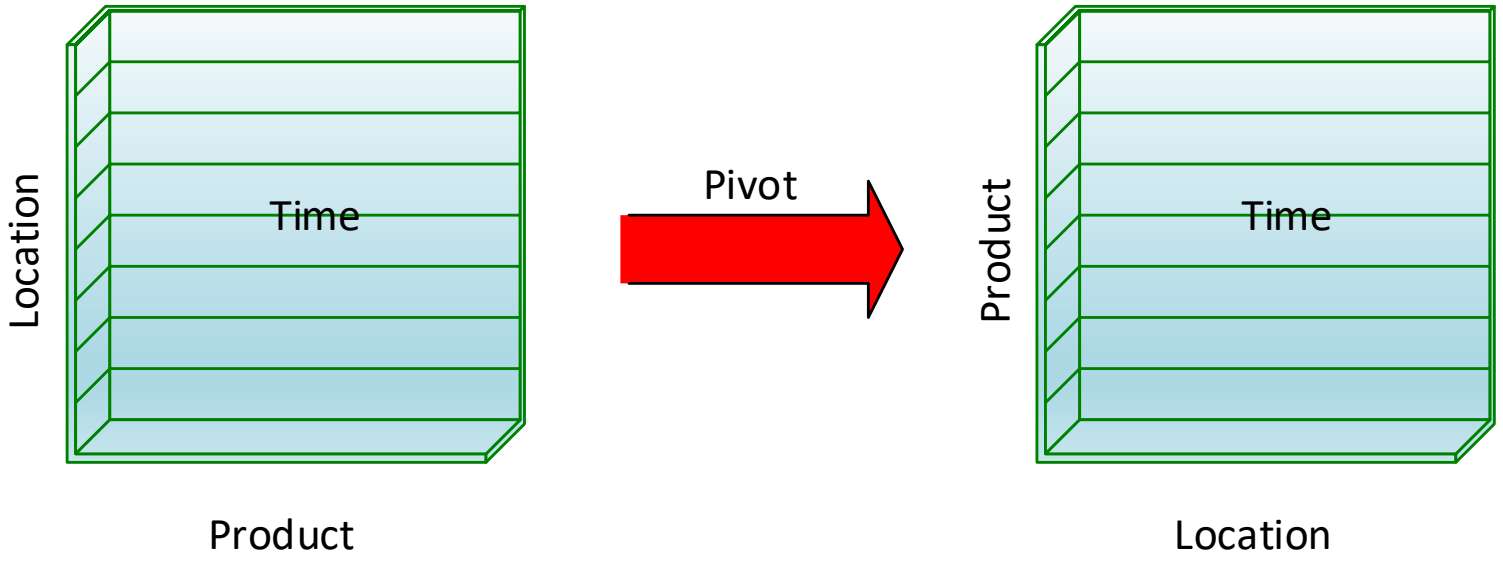
Location	Product			
	Mono Laser	*Ink Jet*	*Photo*	*Portable*
Utah	40	90	50	30
Arizona	70	55	60	35
Colorado	65	45	85	60

Drill-down Example

Location	Product			
	Mono Laser	Ink Jet	Photo	Portable
California	80	110	60	25
- Utah				
Salt Lake	20	20	10	15
Park City	5	30	10	5
Ogden	15	40	30	10
Arizona	70	55	60	35
Washington	75	85	45	45
Colorado	65	45	85	60

Pivot Operator

Rotate or rearrange dimensions



Operator Summary

Operator	Purpose	Description
Slice	Focus attention on a subset of dimensions	Replace a dimension with a single member value or with a summary of its measure values
Dice	Focus attention on a subset of member values	Replace a dimension with a subset of members
Drill-down	Obtain more detail about a dimension	Navigate from a more general level to a more specific level
Roll-up	Summarize details about a dimension	Navigate from a more specific level to a more general level
Pivot	Present data in a different order	Rearrange the dimensions in a data cube

Representation of Data in DW

Dimensional Modeling

- A retrieval-based system that supports high-volume query access

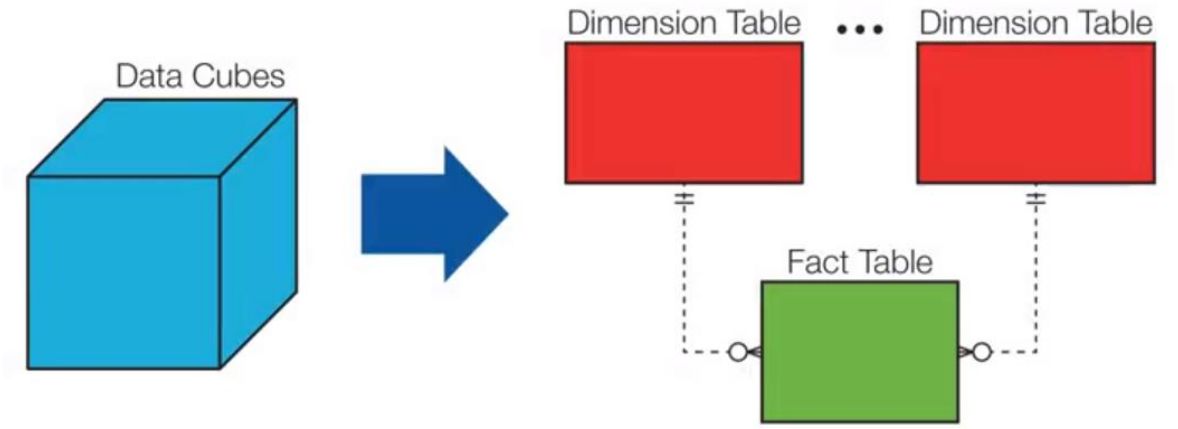
Star schema

- The most commonly used and the simplest style of dimensional modeling
- Contain a **fact table** surrounded by and connected to several **dimension tables**

Snowflakes schema

- An extension of star schema where the diagram resembles a snowflake in shape

Schema design



The term "**schema**" refers to the organization of data

A database uses relational model, while a data warehouse uses Star and/or Snowflake

Multidimensionality

The ability to organize, present, and analyze data by several dimensions, such as sales by region, by product, by salesperson, and by time (four dimensions)

Multidimensional presentation

- **Dimensions:** products, salespeople, market segments, business units, geographical locations, distribution channels, country, or industry
- **Measures:** money, sales volume, head count, inventory profit, actual versus forecast
- **Time:** daily, weekly, monthly, quarterly, or yearly

The “Classic” Star Schema

- A relational model with a one-to-many relationship between dimension table and fact table.
- A single fact table, with detail and summary data
- Fact table primary key has only one key column per dimension
- Each dimension is a single table, highly denormalized

Benefits: Easy to understand, intuitive mapping between the business entities, easy to define hierarchies, reduces # of physical joins, low maintenance, very simple metadata

Drawbacks: Summary data in the fact table yields poorer performance for summary levels, huge dimension tables a problem

Snowflake Schema

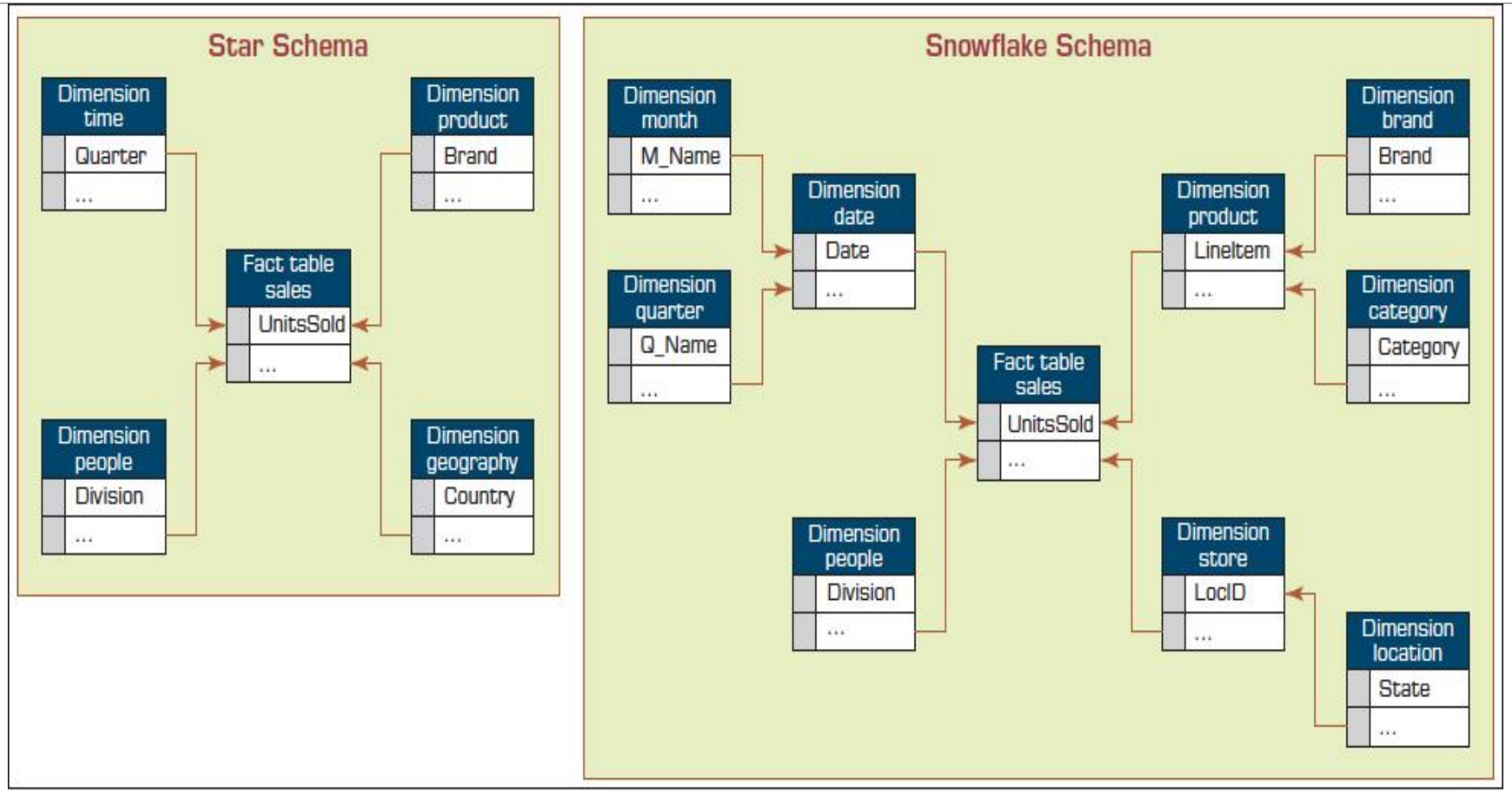
Snowflake schema is a type of star schema but a more complex model.

“Snowflaking” is a method of normalizing the dimension tables in a star schema.

The normalization eliminates redundancy.

The result is more complex queries and reduced query performance.

Star Schema versus Snowflake Schema



DW Implementation: Things to Avoid

Starting with the wrong sponsorship chain

Setting expectations that you cannot meet

Engaging in politically naive behavior

Loading the data warehouse with information just because it is available

Believing that data warehousing database design is the same as transactional database design

Choosing a data warehouse manager who is technology oriented rather than user oriented

DW Implementation: Things to Avoid

Focusing on traditional internal record-oriented data and ignoring the value of external data and of text, images, etc.

Delivering data with confusing definitions

Believing promises of performance, capacity, and scalability

Believing that your problems are over when the data warehouse is up and running

Focusing on ad hoc and periodic reporting instead of alerts

Traditional Data Warehouse Challenges

Organisations need data-driven insights

Traditional DW/BI can't deliver

Need	Support
Unlimited source data breadth and depth	Challenge
Quickly add new data sets	Challenge
Real-time data and analysis	Challenge
Semi-structured and unstructured data	Challenge
Support predictive and discovery analytics	Challenge

Massive DW and Scalability

The main issues pertaining to scalability:

- The amount of data in the warehouse
- How quickly the warehouse is expected to grow
- The number of concurrent users
- The complexity of user queries

- Good scalability means that queries and other data-access functions will grow linearly with the size of the warehouse

The 1990's DW/BI architecture and core technologies are aging!

Goal

Decrease idea-to-insight cycle

Good news – Big Data technology and architecture, e.g. Hadoop and Modern Analytics to the Rescue!

Market Shares and Trends

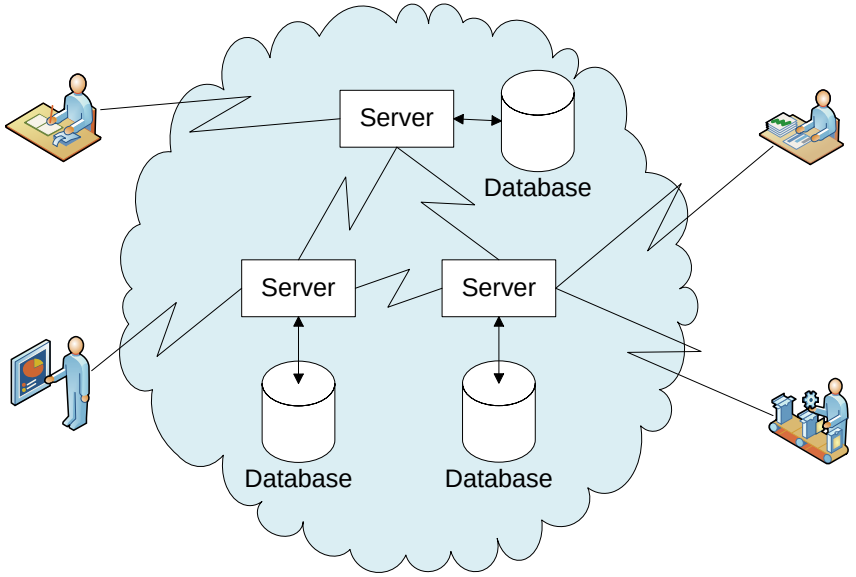
Major DW vendors (survivors!): Teradata, Oracle, IBM, Microsoft, SAP, Pivotal

DW survivors, originally on-premise, now hybrid

Trends

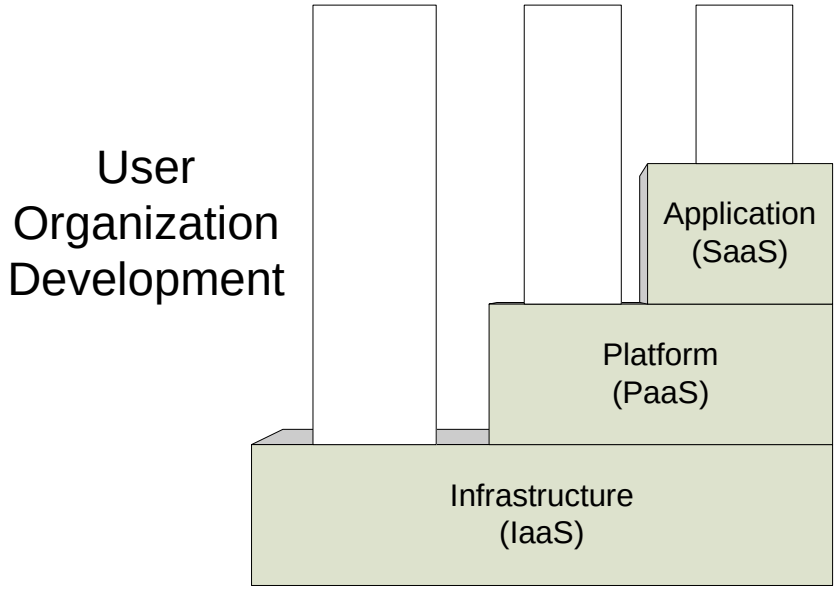
- Real time load and analysis
- Increased storage and analysis of social interactions
- Increased usage of cloud services and appliances

Cloud Influence



- Reduces local expertise to procure technology and manage a data warehouse
- Economies of scale
- Improved scalability
- Higher variable costs but lower fixed costs

Cloud Service Models



The Future of DW

Sourcing…

- *Web, social media, and Big Data*
- Open source software
- SaaS (software as a service)
- Cloud computing
- Data lakes

Infrastructure…

- Columnar
- Real-time DW
- Data warehouse appliances
- Data management practices/technologies
- In-database & In-memory processing New DBMS
- New DBMS, Advanced analytics, …

Resources:

Data Warehousing Concepts: a brief overview of several concepts in data warehousing. It describes dimensional modeling and some of its features and different types of OLAP technology.

Inmon vs Kimball – which approach?

Star vs snowflake – which is better?

Sandboxes – e-bay's example - user controlled spaces for data analysis