Electronics and
Computer Science

UNIVERSITY OF
Southampton

Data Warehousing

COMP3017 Advanced Databases

Dr Nicholas Gibbins – nmg@ecs.soton.ac.uk
2013-2014

UNIVERSITY OF Southampton

Processing Styles – OLTP

On-Line Transaction Processing

- Traditional workloads, ‘bread and butter’ processing
- Volumes of data, transactions grow, networks getting larger.

UNIVERSITY OF Southampton

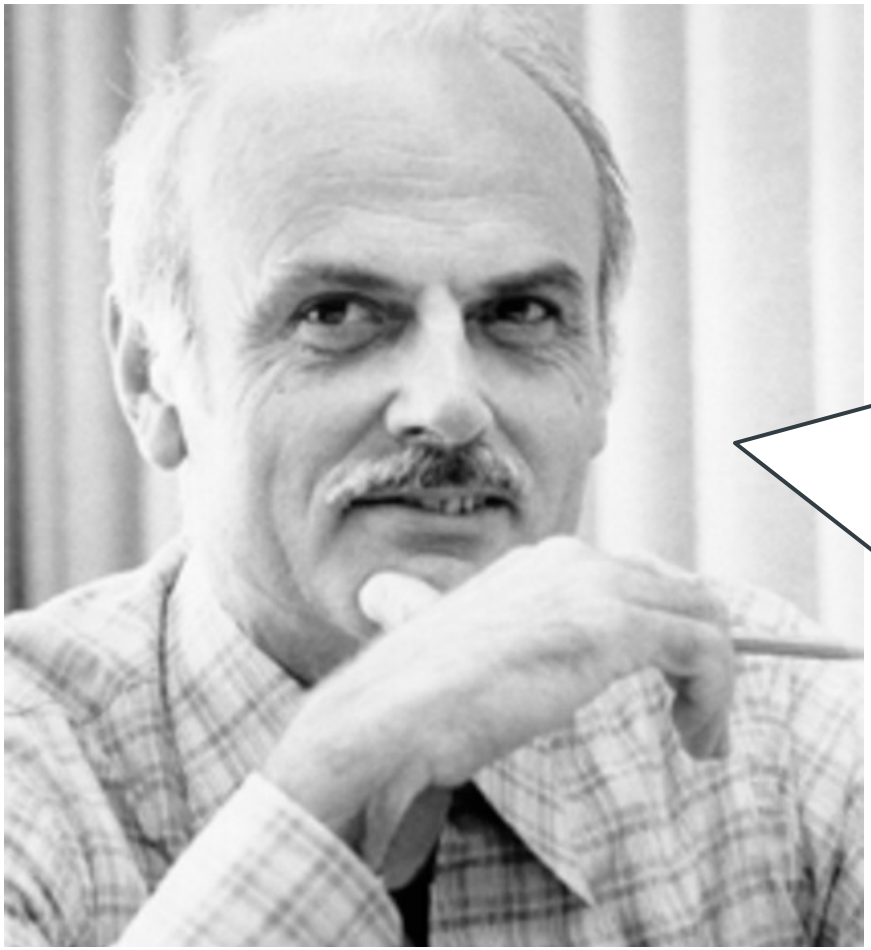
Processing Styles – OLAP

On-Line Analytical Processing

- includes the use of data warehouses
- multidimensional databases
- data analysis

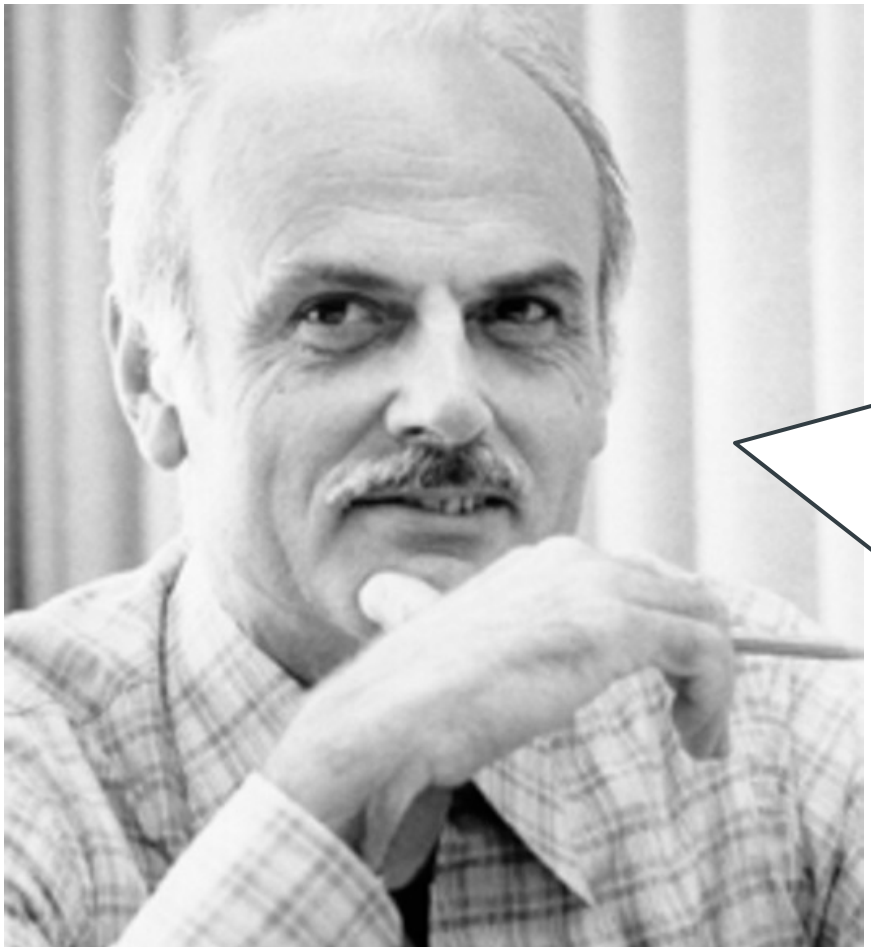
UNIVERSITY OF Southampton

Online Analytical Processing



OLAP is the name given to the dynamic enterprise analysis required to create, manipulate, animate and synthesise information from exegetical, contemplative and formulaic data analysis models

Online Analytical Processing

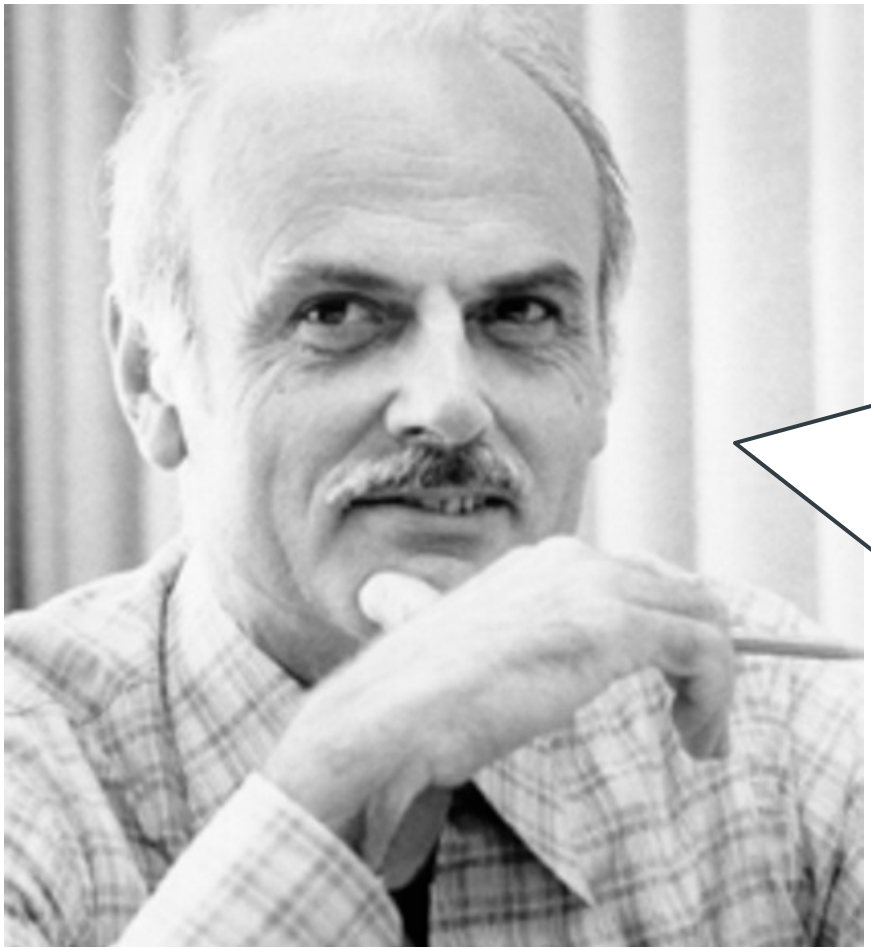


OLAP is the name given to the dynamic enterprise analysis required to create, manipulate, animate and synthesise information from **exegetical**, contemplative and formulaic data analysis models

Exegesis: critical explanation

How did we get to where we are?

Online Analytical Processing

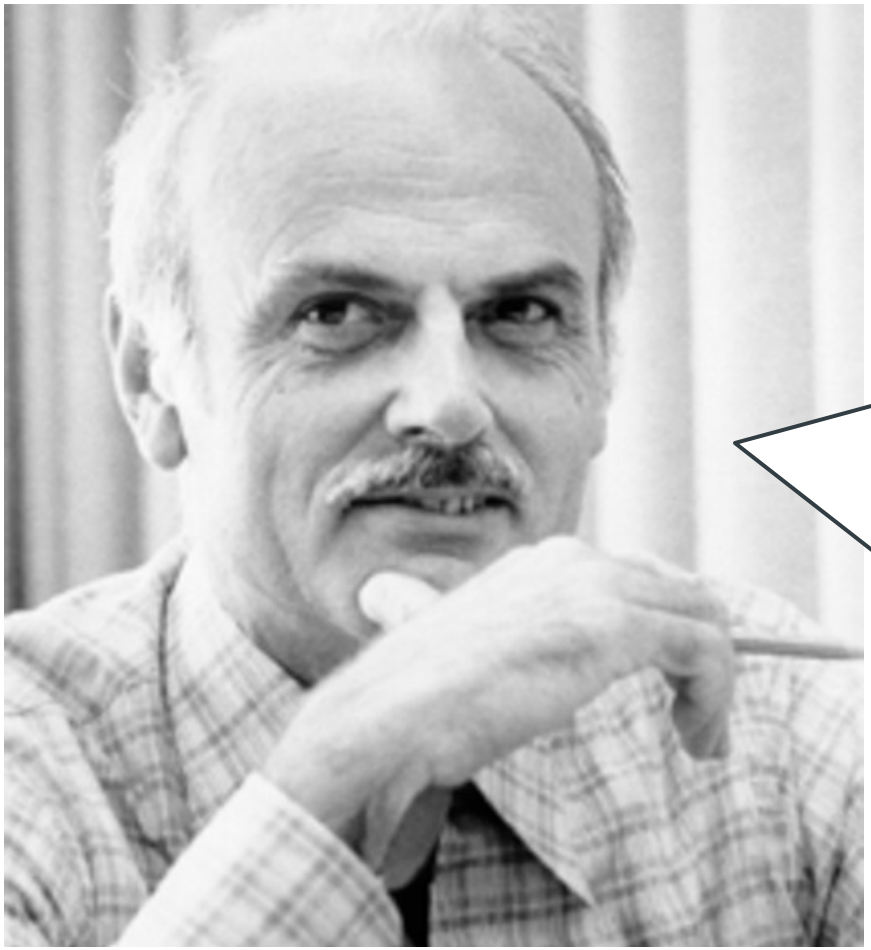


OLAP is the name given to the dynamic enterprise analysis required to create, manipulate, animate and synthesise information from exegetical, **contemplative** and formulaic data analysis models

Asking ‘what if?’ questions

How does the outcome change if we vary the parameters?

Online Analytical Processing



OLAP is the name given to the dynamic enterprise analysis required to create, manipulate, animate and synthesise information from exegetical, contemplative and **formulaic** data analysis models

Which parameters must be varied in order to achieve a given outcome?

University of Southampton

12 Rules for OLAP

1. Multidimensional conceptual view
2. Transparency
3. Accessibility
4. Consistent reporting performance
5. Client-server architecture
6. Generic dimensionality
7. Dynamic sparse matrix handling
8. Multi-user support
9. Unrestricted cross-dimensional operations
10. Intuitive data manipulation
11. Flexible reporting
12. Unlimited dimensions and aggregation levels

Data Mining

- *Data mining* is the process of discovering hidden patterns and relations in large databases using a variety of advanced analytical techniques
- Data mining attempts to use the computer to discover relationships that can be used to make predictions
- Data mining tools often find unsuspected relationships in data that other techniques will overlook

UNIVERSITY OF Southampton

Data Mining Approaches

- Rule-based analysis
- Neural networks
- Fuzzy Logic
- K-nearest-neighbour
- Genetic algorithms
- Advanced visualisation
- Combination of any of the above

UNIVERSITY OF Southampton

The Data Warehouse

A *data warehouse* is a subject-oriented, integrated, time-variant, non-volatile collection of data that is used primarily in organisational decision making

The Data Warehouse

A *data warehouse* is a **subject-oriented**, integrated, time-variant, non-volatile collection of data that is used primarily in organisational decision making

The data is organised according to subject instead of application and contains only the information necessary for ‘decision support’ processing.

UNIVERSITY OF Southampton

The Data Warehouse

A *data warehouse* is a subject-oriented, **integrated**, time-variant, non-volatile collection of data that is used primarily in organisational decision making

Data encoding is made uniform
(e.g. sex = f or m, 1 or 2, b or g - needs to be all the same in the warehouse).

Data naming is made consistent.

The Data Warehouse

A *data warehouse* is a subject-oriented, integrated, **time-variant**, non-volatile collection of data that is used primarily in organisational decision making

Data is collected over time and can then be used for comparisons, trends and forecasting

UNIVERSITY OF Southampton

The Data Warehouse

A *data warehouse* is a subject-oriented, integrated, time-variant, **non-volatile** collection of data that is used primarily in organisational decision making

The data is not updated or changed once in the data warehouse, but is simply loaded, and then accessed.

The data warehouse is held quite separately from the operational database, which supports OLTP.

University of Southampton

Why a Separate Data Warehouse?

Performance

- Operational databases are optimised to support known transactions and workloads
- Special data organisation, access methods and implementation methods are needed
- Complex OLAP queries would degrade performance for operational transactions

Why a Separate Data Warehouse?

Missing data

- Decision support requires historical data, which operational databases do not typically maintain

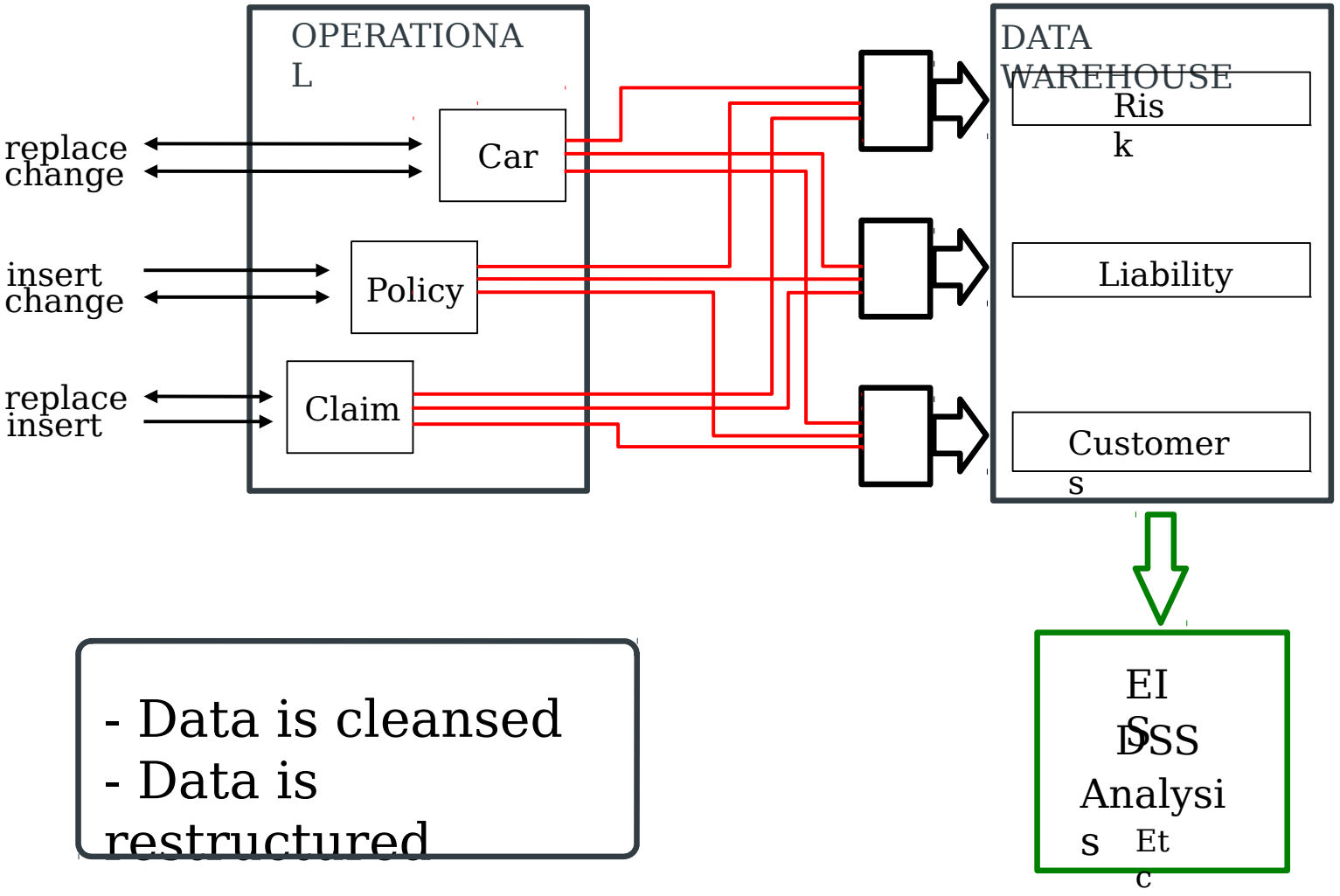
Data consolidation

- Decision support requires consolidation (aggregation, summarisation) of data from many heterogeneous sources, including operational databases and external sources

Data quality

- Different sources typically use inconsistent data representations, codes and formats, which have to be reconciled

Extracting Data



- Data is cleansed
- Data is restructured

The Data Warehouse

A Data Warehouse may be realised:

- via a front end to existing databases and files
- in a fresh relational database
- in a multidimensional database (MDDB)
- in a proprietary database format
- using a mixture of the above

University of Southampton

The Data Warehouse

Data may be accessed in various ways:

- Decision Support Systems (DSS)
- Executive Information Systems (EIS)
- Data Mining
- On-Line Analytical Processing

UNIVERSITY OF Southampton

Data Marts

- A data mart focuses on
  - only one subject area, or
  - only one group of users
- An organisation can have
  - one enterprise data warehouse
  - many data marts
- Data marts do not contain operational data
- Data marts are more easily understood and navigated

UNIVERSITY OF Southampton

Multidimensional Analysis

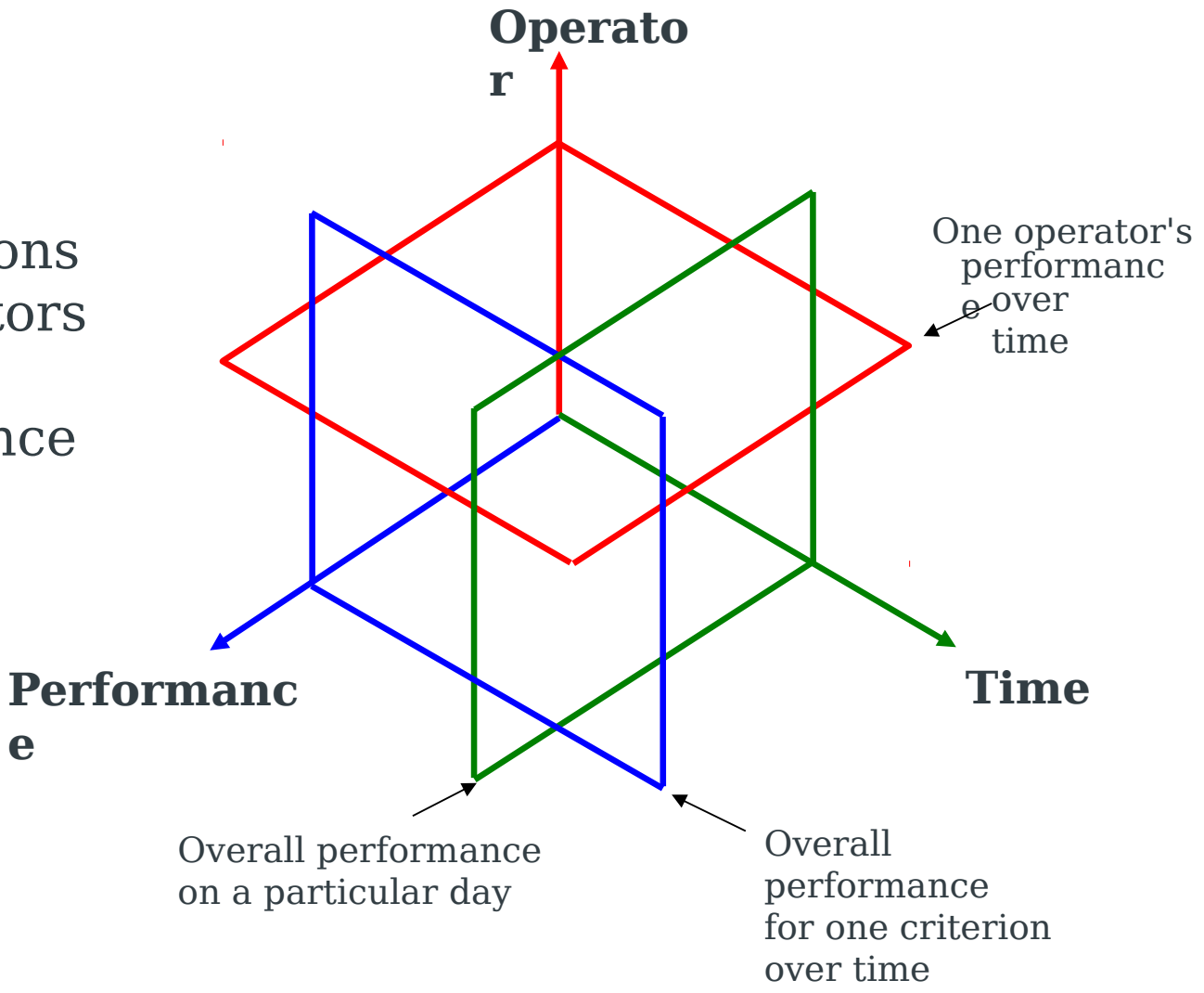
Need to examine data in various ways

Produce views of multidimensional data for users:

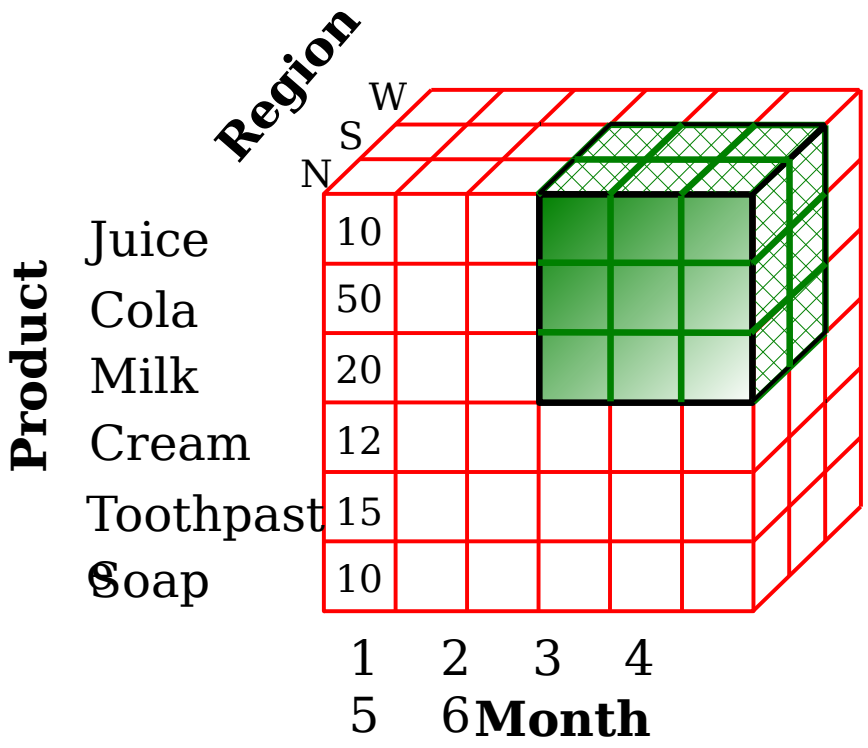
- Slice
- Dice
- Pivot
- Drill down
- Roll up

Multidimensional Analysis – Slice

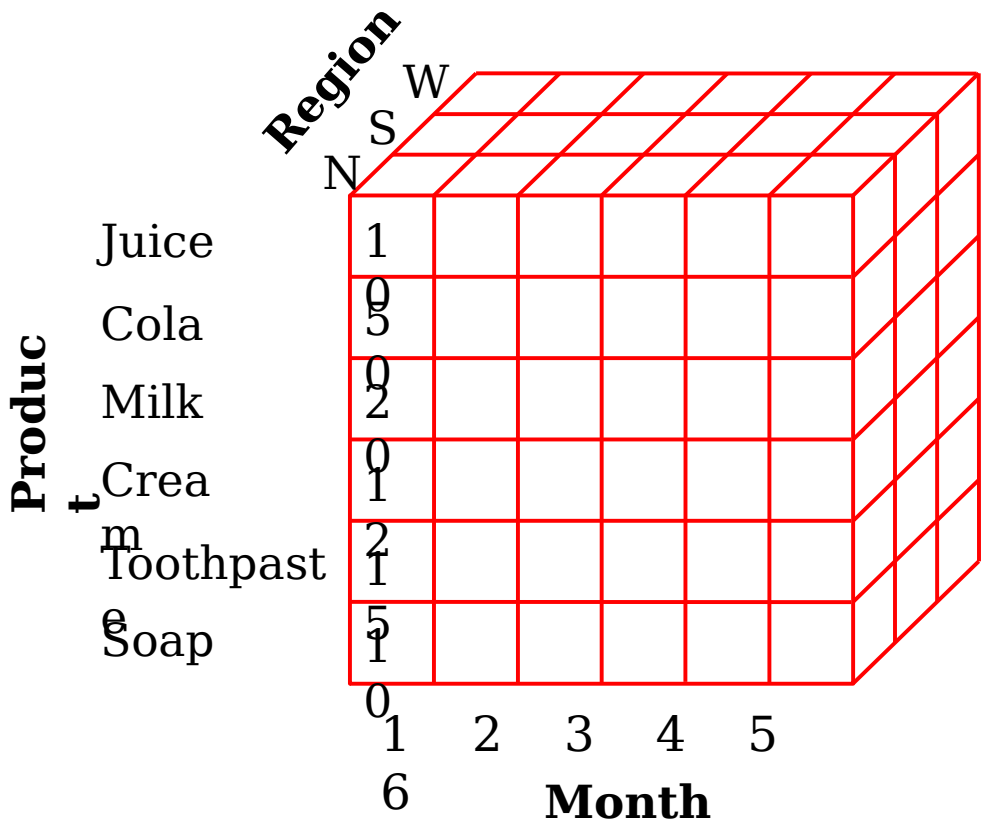
Train Performance
- 3 dimensions
  - Operators
  - Performance
  - Time

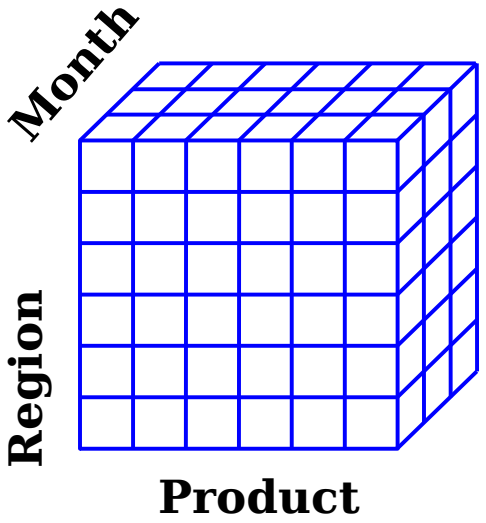


Multidimensional Analysis – Dice

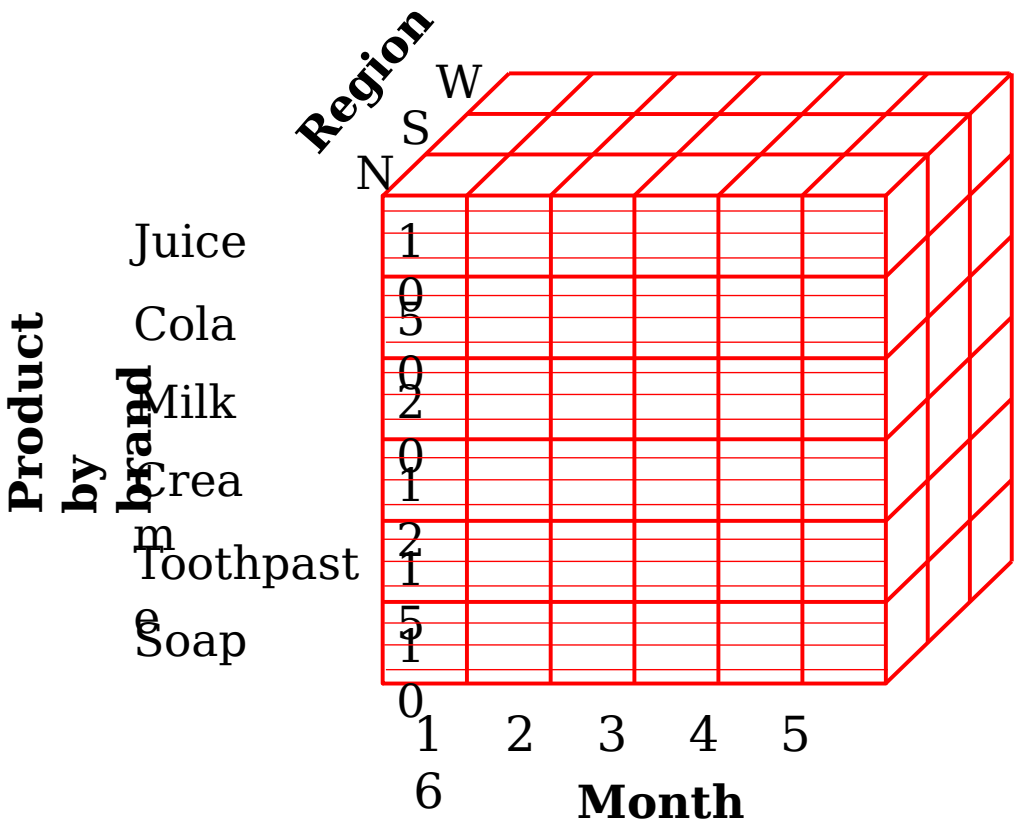


Multidimensional Analysis – Pivot

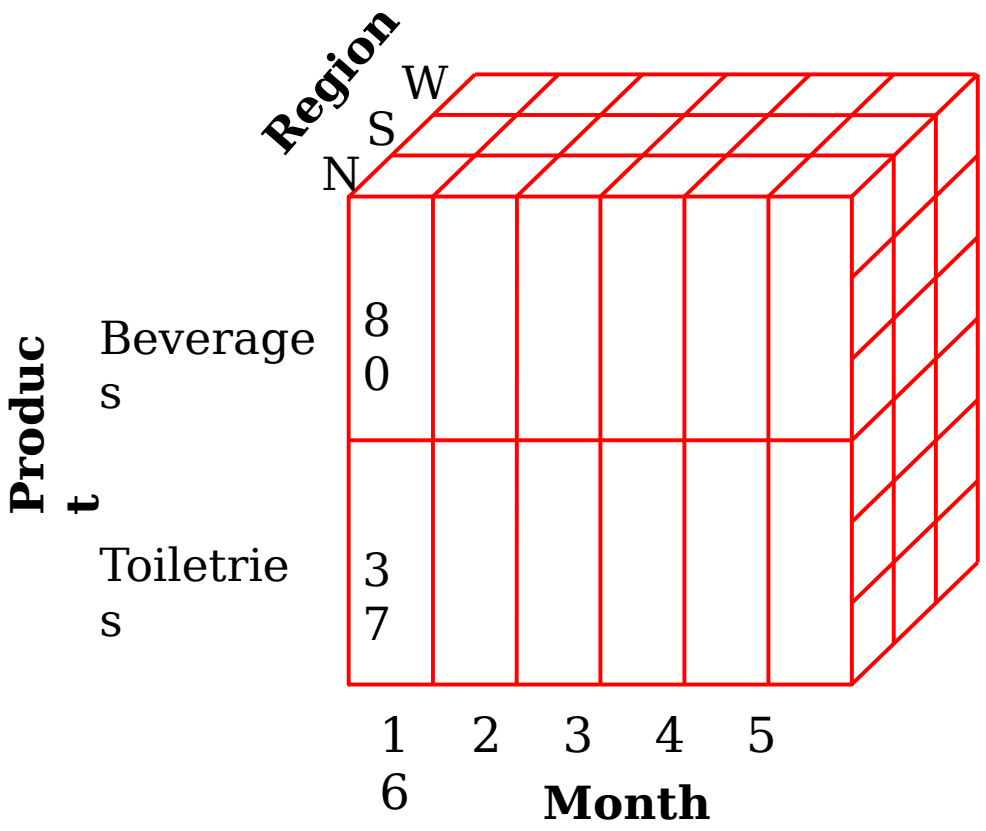




Multidimensional Analysis – Drill Down



Multidimensional Analysis – Roll Up



UNIVERSITY OF Southampton

Internal Aspects

Schemas

- Star schema
- Snowflake schema
- Fact constellation schema

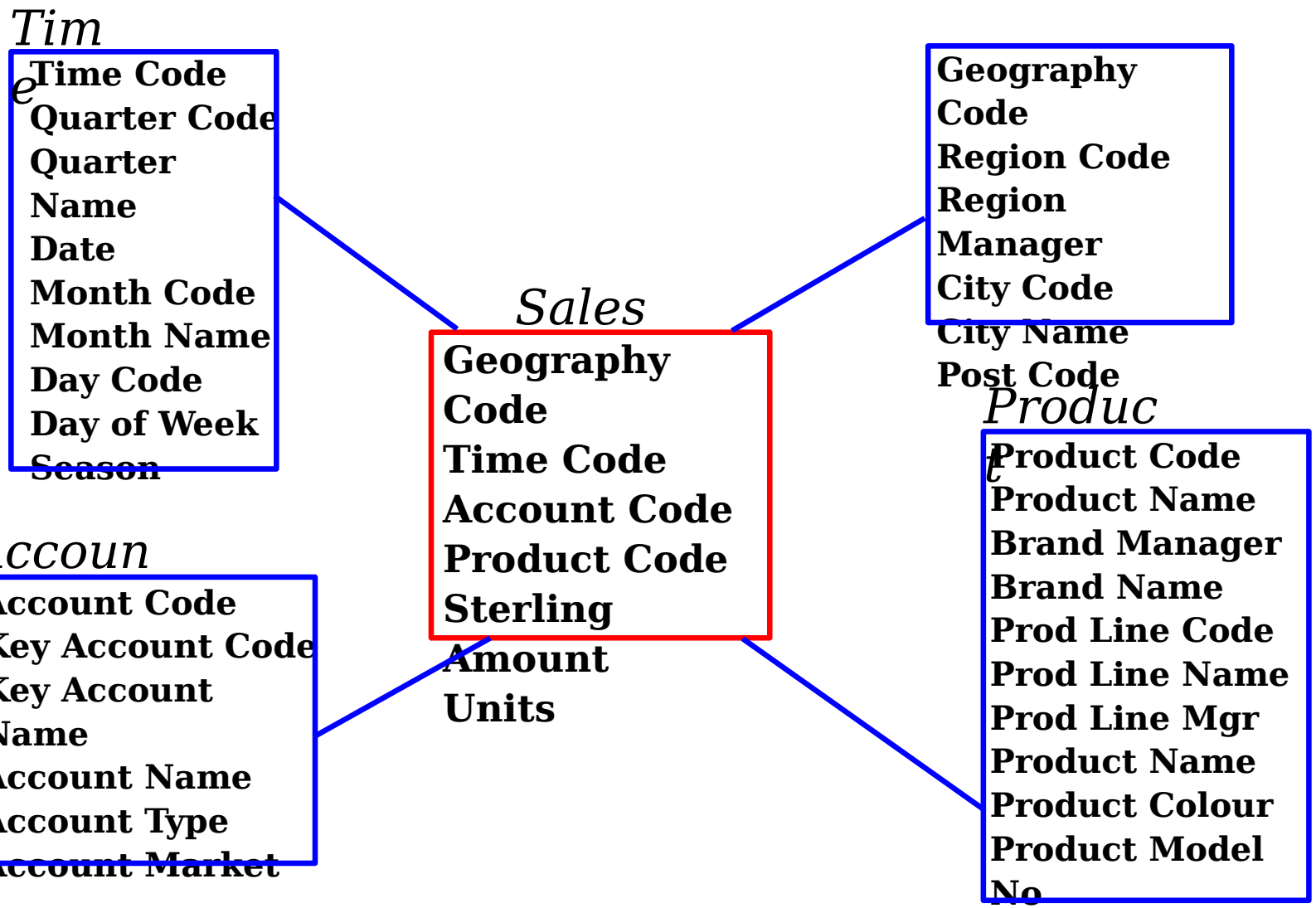
Aggregated data

Specialised indexes

- Bit map indexes (see lecture on multidimensional indexes)
- Join indexes

Specialised join methods

Star Schema


Time
Time Code
Quarter Code
Quarter
Name
Date
Month Code
Month Name
Day Code
Day of Week
Season
Account
Account Code
Key Account Code
Key Account Name
Account Name
Account Type
Account Market
Sales
Geography Code
Time Code
Account Code
Product Code
Sterling Amount
Units
Geography
Code
Region Code
Region
Manager
City Code
City Name
Post Code
Product
Product Code
Product Name
Brand Manager
Brand Name
Prod Line Code
Prod Line Name
Prod Line Mgr
Product Name
Product Colour
Product Model No

UNIVERSITY OF Southampton

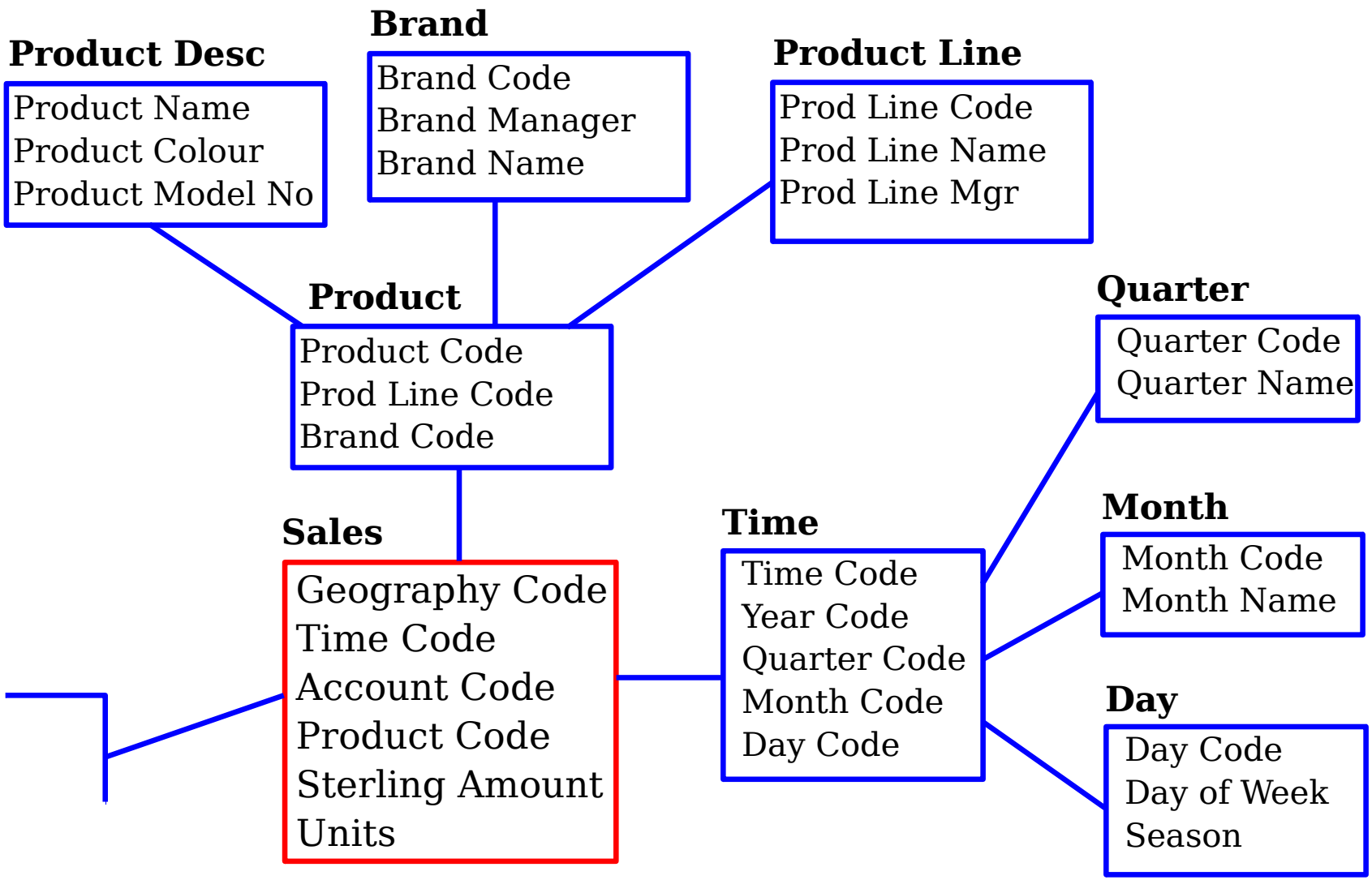
Fact Tables

Prod_Code	Time_Code	Acct_Code	Sales	Qty
101	2045	501	100	1
102	2045	501	225	2
103	2046	501	200	20
104	2046	502	250	25
105	2046	502	20	1

key columns joining fact table to the dimension tables

numerical measures

Part of a Snowflake Schema



UNIVERSITY OF Southampton

Data Warehouse Databases

Relational and Specialised RDBMSs

- Specialised indexing techniques, join and scan methods

Relational OLAP (ROLAP) servers

- Explicitly developed to use a relational engine to support OLAP
- Include aggregation navigation logic, the ability to generate multi-statement SQL, and other additional services

Multidimensional OLAP (MOLAP) servers

- The storage model is an n-dimensional array
- May use a 2-level approach, with 2-D dense arrays indexed by B-Trees