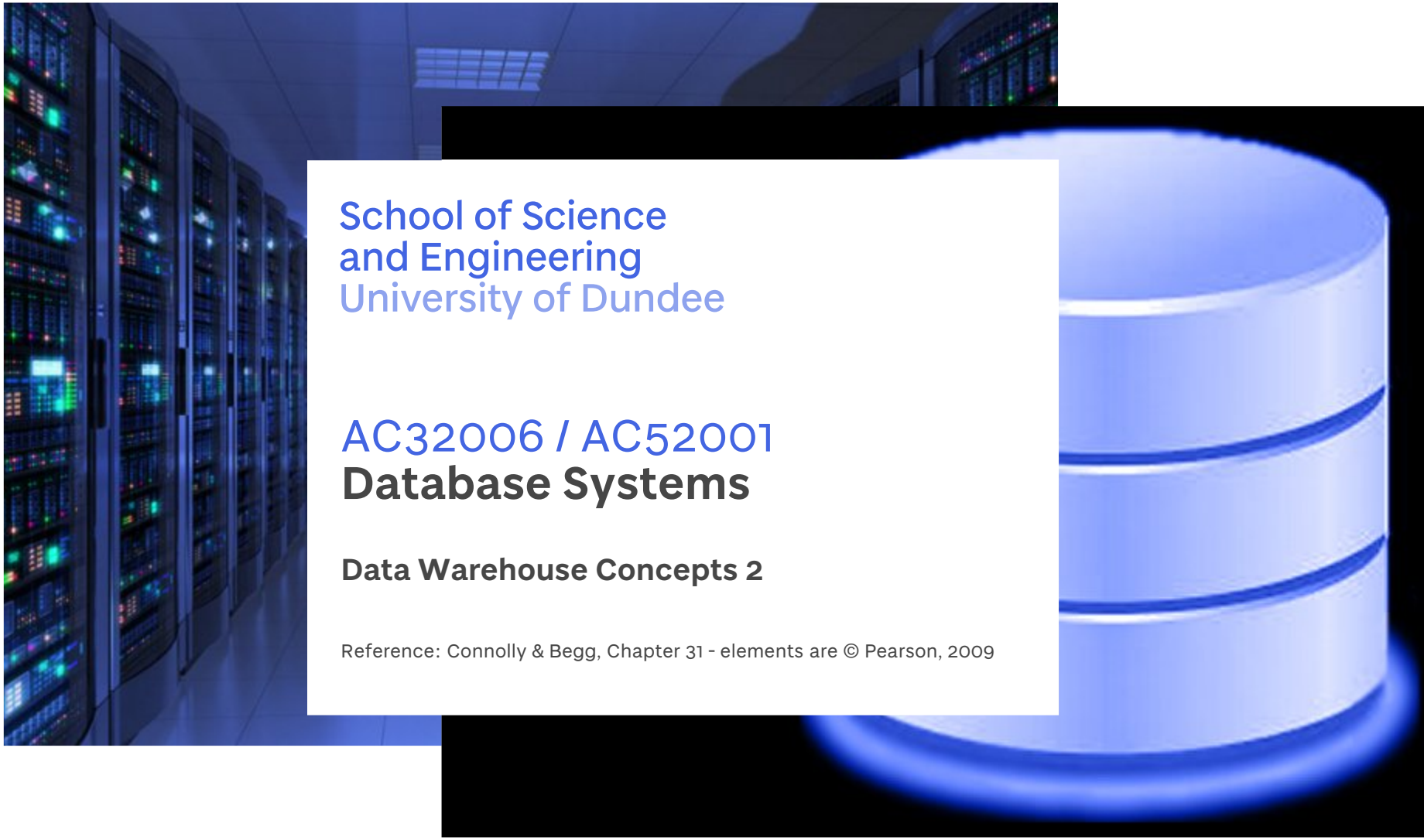


School of Science
and Engineering
University of Dundee

AC32006 / AC52001
Database Systems

Data Warehouse Concepts 2

Reference: Connolly & Begg, Chapter 31 - elements are © Pearson, 2009

In this video ...

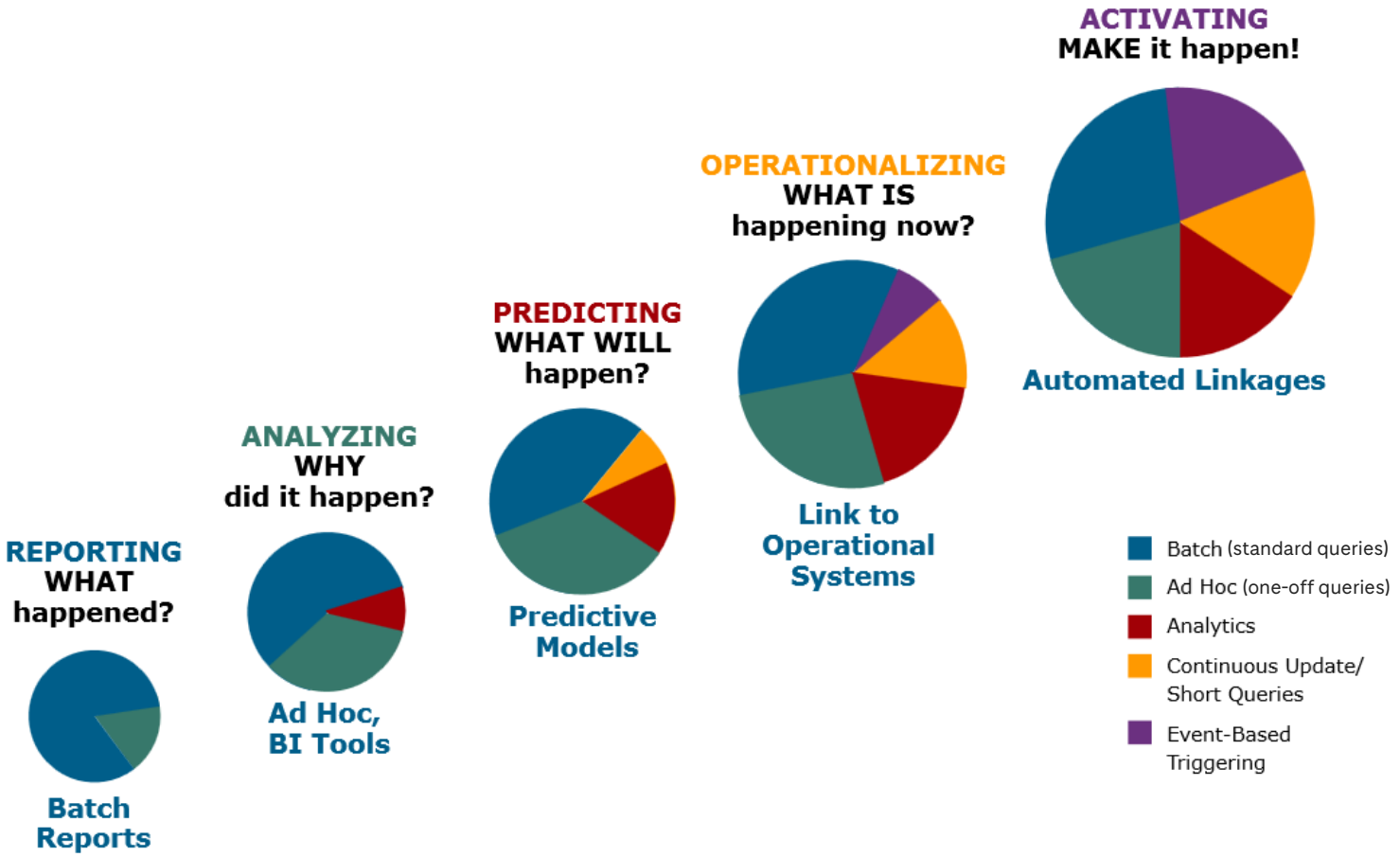
→the reasons for data warehousing

→the main features of a data warehouse

Revision - Data Warehousing

- We have seen the evolution of data warehouses as a time-variant but non-volatile collection of data from an enterprise over a period of time (rather than at an *instant* in time)
- A DW is designed to permit sophisticated queries to be run which offer the ability to predict the future, potentially leading to cost savings (or other benefits) for the enterprise

Data Warehouse Evolution



"Five Stages of an Active Data Warehouse Evolution", Stephen Brobst and Joe Rarey, *Teradata Magazine* Online

OLTP vs. Data Warehousing

Characteristic	OLTP	Data Warehouse
Main purpose	Support operational processes	Support analytical processes
Data age	Current	Historic (trend is towards RT)
Data latency	Real-time	Depends on length of cycle for data (trend is towards RT)
Data granularity	Detailed data	Detailed data, also lightly and highly summarised data
Data processing	Predictable pattern of insertions, deletions and queries; high transaction throughput	Less predictable pattern of queries; low-to-medium transaction throughput
Reporting	Predictable, one-dimensional, relatively static fixed reporting	Unpredictable multi-dimensional dynamic reporting
Users	Large number of operational users	Low number of managerial users (trend is towards analytical requirements of operational users)

Benefits of Data Warehousing

- → Potential high returns on investment
- → Competitive advantage
- → Increased productivity of corporate decision-makers
- → Leveraging the data available in OLTP systems

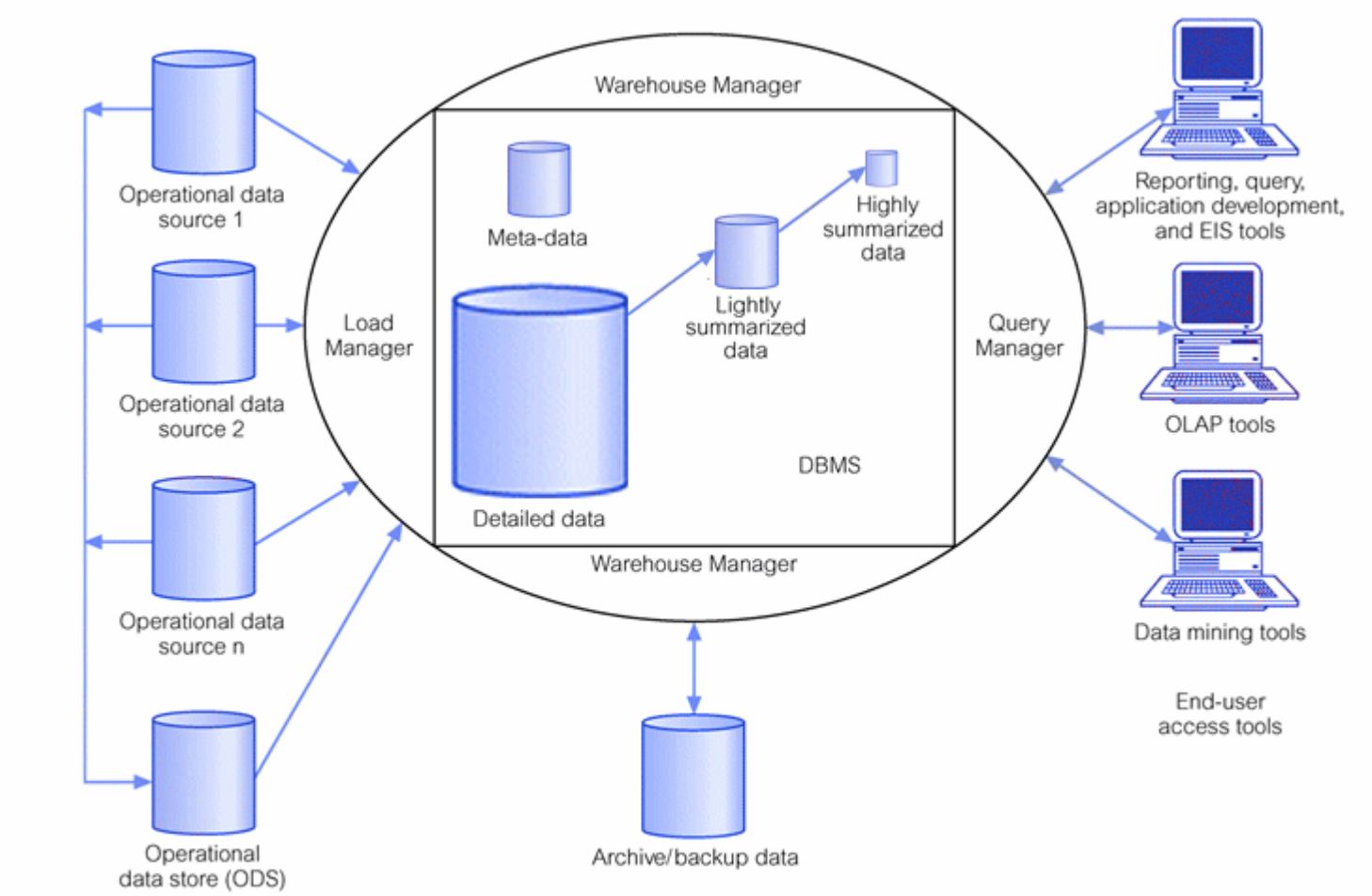
General Problems of Data Warehousing

- Handling data from lots of different sources and different formats
- People:
  - have *same* terminology for *different* things
  - have *different* terminology for *same* things
  - like ownership of “their” data
  - want different things
- Cost - complex design and huge storage requirements … but *should* be offset by the benefits offered

Specific Problems of Data Warehousing

- → Underestimation of resources for data loading:
  - → looks simpler than it really is (*Here Be Dragons!*)
  - → hidden problems with source systems
  - → required data has not been captured
  - → complexity of integration and data homogenisation
- → Increased end-user demands (scope creep)
- → High demand for resources
- → Data ownership
- → High maintenance
- → Long duration projects

Example Data Warehouse Architecture



(typically OLTP databases, spreadsheets, etc.)

Operational Data Sources

- The *main* sources of data are online transaction processing (OLTP) databases; typically an organisation will already have these (but they may be *incompatible*)
- Useful to include *any other relevant sources* such as personal databases and spreadsheets (which management *may* not know about), Enterprise Resource Planning (ERP) files, and web usage log files

Operational Data Store (ODS)

- → Holds current and integrated operational data for analysis
- → Often structured and supplied with data in the same way as the data warehouse
- → May act as a staging area for data to be moved into the warehouse
- → Often created when legacy operational systems are found to be incapable of achieving reporting requirements

Load Manager / ETL Manager

- In most cases, the data sources are in different formats, inconsistent and contain "dirty" or incomplete data
- We cannot put this directly into the data warehouse – it must be cleaned and re-organised into a more usable format
- This process is called Extract Transform and Load (ETL)
- ETL typically accounts for ~75% of the work in setting up a data warehouse

Load Manager / ETL Manager

- Data for a data warehouse must be extracted from one or more data sources, transformed into a form that is easy to analyse and is consistent with data already in the warehouse, and *then* finally loaded into the DW
- There are tools which automate the ETL processes and also offer additional facilities such as data profiling, data quality control, and metadata management

ETL Processes

Extraction:

- → Targets one or more data sources:
  - → OLTP databases (typically), but also …
  - → personal databases
  - → personal spreadsheets
  - → Enterprise Resource Planning (ERP) files
  - → web usage log files
- → The data sources are normally *internal* but can also include *external* sources such as the systems used by suppliers, customers and others e.g. weather data

ETL Processes

Transformation:

→ Applies a series of rules or functions to the extracted data, which determines how the data will be used for analysis

→ Can involve transformations such as data summations, data encoding, data merging, data splitting, data calculations, and creation of surrogate keys

ETL Processes

Loading:

→ As data loads, additional constraints defined in the database schema can be activated (such as uniqueness, referential integrity, and mandatory fields), which contribute to the overall data quality performance of the ETL process

ETL Tools

Data profiling and *quality control*:

→ Provides important information about the quantity and quality of the data coming from the source systems

Metadata management:

→ Understanding a query result can require consideration of the data history i.e. what happened to the data during the ETL process? The answers are held in the metadata repository.

Warehouse Manager

Performs all of the DBMS functions associated with the operational management of the data in the warehouse, such as:

→ analysis of data to ensure consistency

→ transformation and merging of source data from temporary storage into data warehouse tables

→ creation of indexes and views on base tables

→ generation of de-normalisations (if necessary)

→ generation of aggregations (if necessary)

→ backing-up and archiving data

Query Manager

- Performs the operations associated with the management of user queries such as:
  - directing queries to the appropriate tables and scheduling the execution of queries
  - in some cases, the query manager also generates query profiles to allow the warehouse manager to determine which indexes and aggregations are appropriate
    - a query profile can be generated for each user, group of users, or the data warehouse and is based on information that describes the characteristics of the queries such as frequency, target table(s), and size of the results set

End-User Access Tools

- Main purpose of a DW is to *support decision makers* and this is achieved through the provision of a range of access tools including:
  - reporting and querying
  - application and development
  - OLAP
  - data mining

Data Warehouse DBMS Requirements

→ Load performance
→ Load processing
→ Data quality management
→ Query performance
→ Terabyte scalability
→ Mass user scalability
→ Networked data warehouse
→ Warehouse administration
→ Integrated dimensional analysis
→ Advanced query functionality

Some numbers

→ 1 KB = 1,024 bytes
→ 1 MB = $1{,}024^2$ = 1,048,576 bytes
→ 1 GB = $1{,}024^3$ = 1,073,741,824 bytes
→ 1 TB = $1{,}024^4$ = 1,099,511,627,776 bytes
→ 1 PB = $1{,}024^5$ = 1,125,899,906,842,624 bytes

Today's largest data warehouses store *over 50PB* of data – this is *per company/organisation*

i.e. >50,000 times the storage of a modern PC

Source: Teradata Labs

Some numbers

A large data warehouse may be processing:

Largest table?

→ 150 billion rows

How many tables?

→ Over 300,000

Inserts/Updates per day?

→ 3 billion records

Users?

→ Over 300,000

Queries per day?

→ Over 4 million

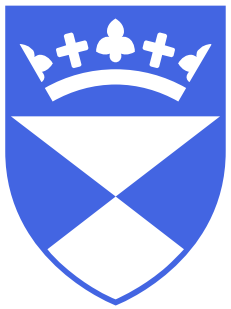
Source: Teradata Labs

Summary

We have seen:

→the reasons for data warehousing

→the main features of a data warehouse



University of Dundee

irmurray@dundee.ac.uk