Ulster University

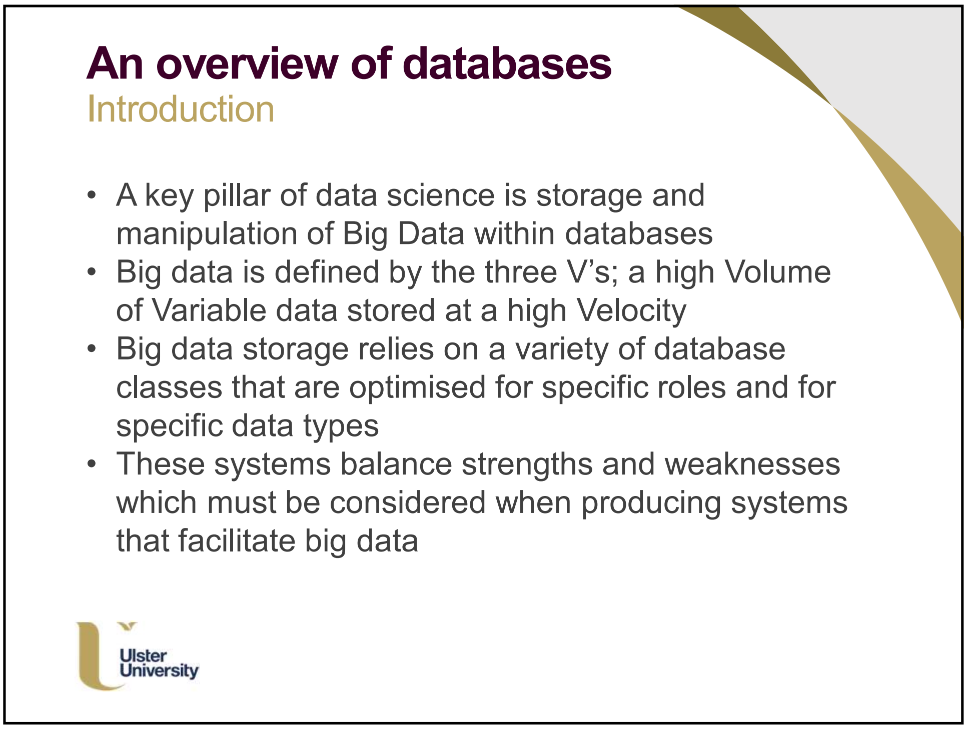
COM 745

Databases within big data

Dr. Joe Rafferty

ulster.ac.uk

1



An overview of databases

Introduction

- A key pillar of data science is storage and manipulation of Big Data within databases
- Big data is defined by the three V's; a high Volume of Variable data stored at a high Velocity
- Big data storage relies on a variety of database classes that are optimised for specific roles and for specific data types
- These systems balance strengths and weaknesses which must be considered when producing systems that facilitate big data

Ulster University

2

An overview of databases

Introduction

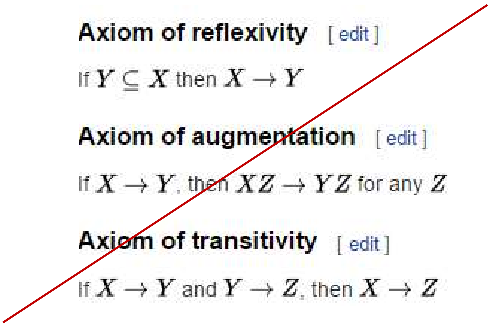
- This module aims to introduce an overview of database systems
- A variety of database systems will be covered
  - The underlying concepts behind these databases will be summarised
  - The optimisations each of these systems will be covered
  - Use cases of these systems will be explored
  - Practical skills related to use of these systems will be explained and taught

3

An overview of databases

Introduction

- This module will not explore implementation of these databases systems



Axiom of reflexivity [edit]

If $Y \subseteq X$ then $X \to Y$

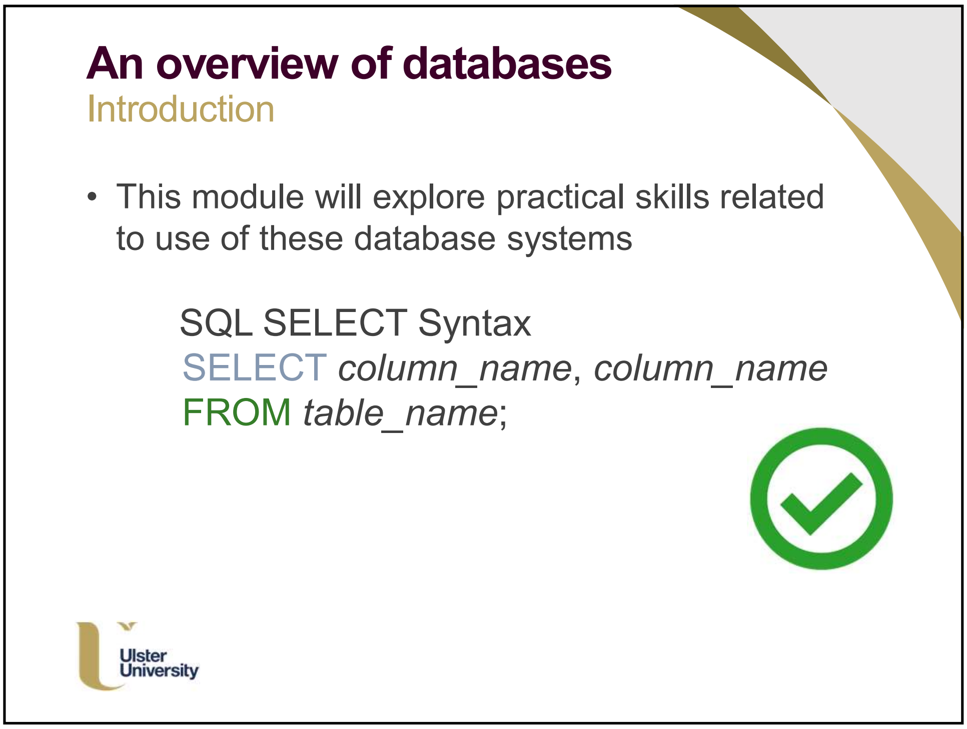
Axiom of augmentation [edit]

If $X \to Y$, then $XZ \to YZ$ for any Z

Axiom of transitivity [edit]

If $X \to Y$ and $Y \to Z$, then $X \to Z$

4



An overview of databases

Introduction

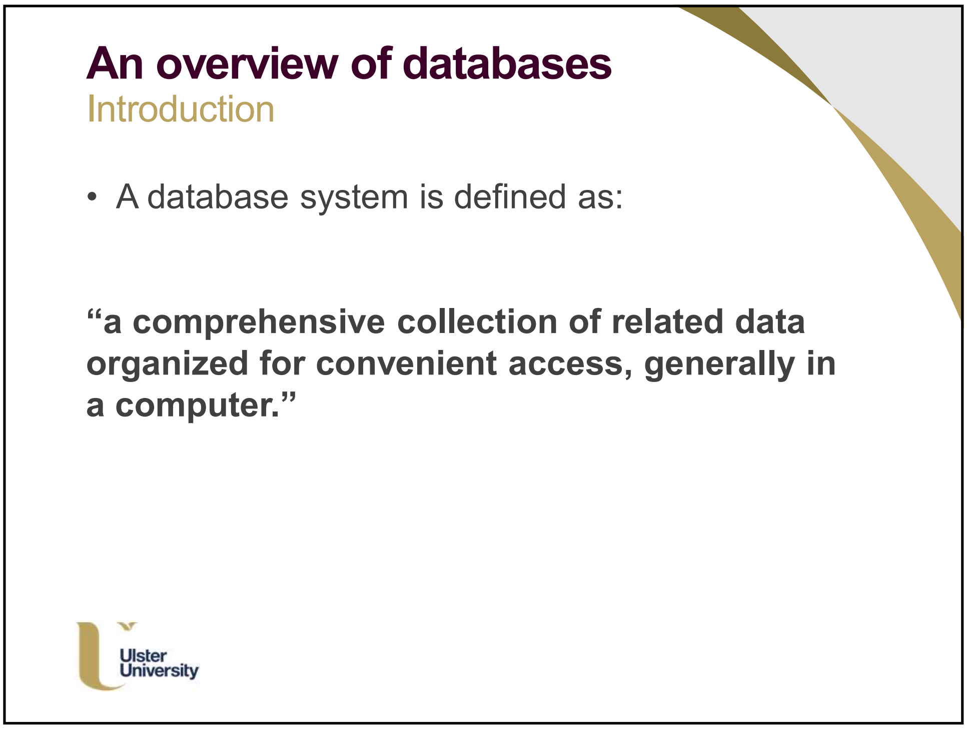
- This module will explore practical skills related to use of these database systems

```
SQL SELECT Syntax
SELECT column_name, column_name
FROM table_name;
```

✓

Ulster University

5



An overview of databases

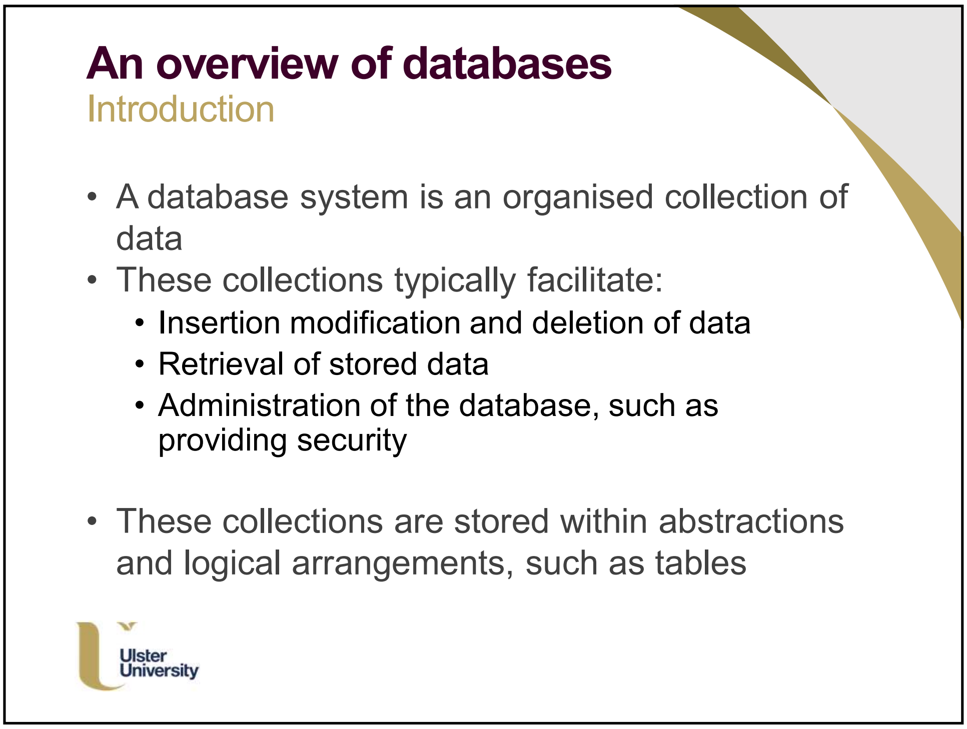
Introduction

- A database system is defined as:

“a comprehensive collection of related data organized for convenient access, generally in a computer.”

Ulster University

6



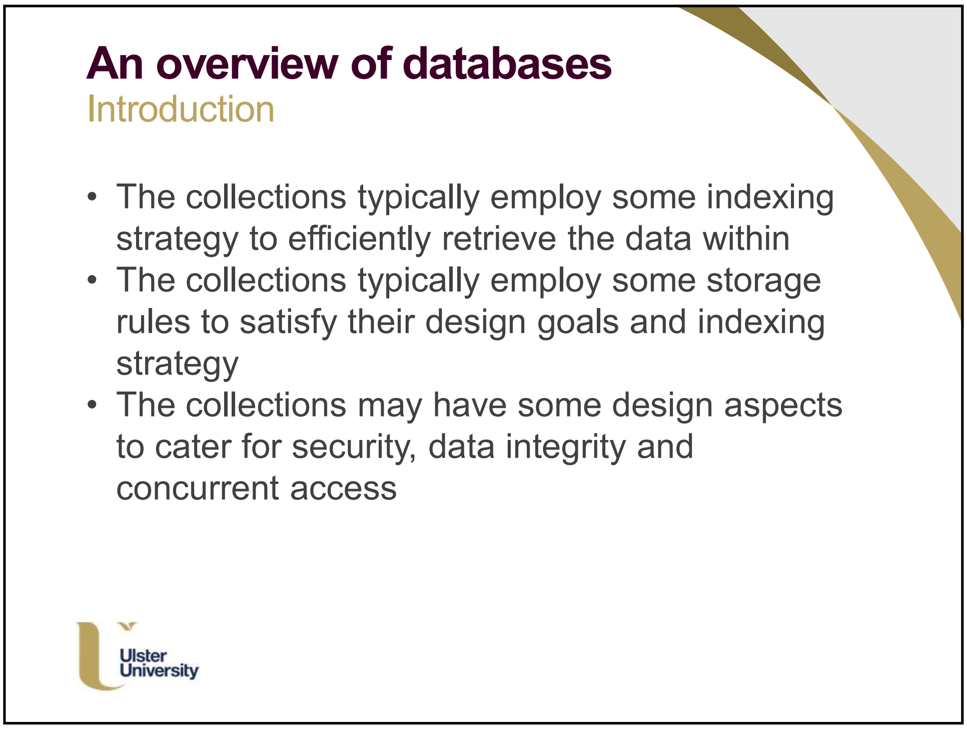
An overview of databases

Introduction

- A database system is an organised collection of data
- These collections typically facilitate:
  - Insertion modification and deletion of data
  - Retrieval of stored data
  - Administration of the database, such as providing security
- These collections are stored within abstractions and logical arrangements, such as tables

Ulster University

7



An overview of databases

Introduction

- The collections typically employ some indexing strategy to efficiently retrieve the data within
- The collections typically employ some storage rules to satisfy their design goals and indexing strategy
- The collections may have some design aspects to cater for security, data integrity and concurrent access

Ulster University

8

An overview of databases

Introduction

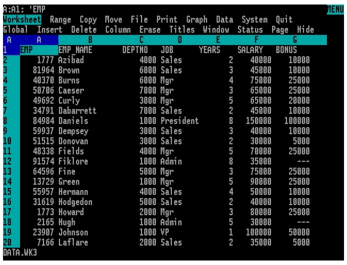
- Databases can reside in the physical world and within computing
- Physical world databases include:
  - An address book
  - A phone book
  - Index Cards
  - The Dewey Decimal System*

Ulster University

An overview of databases

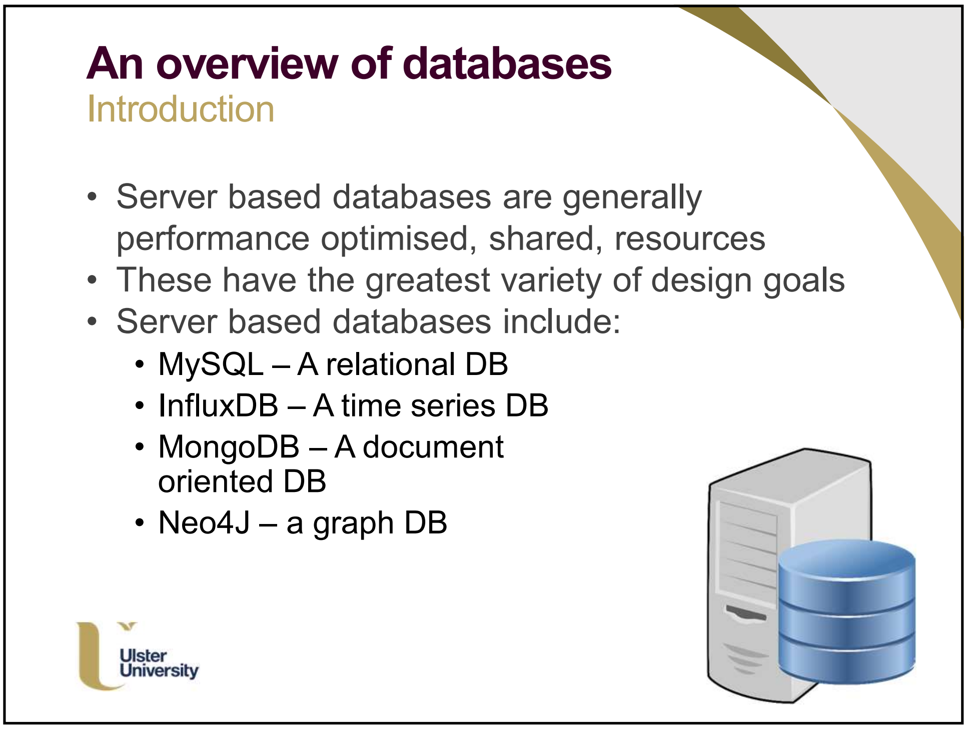
Introduction

- Generally computing based databases can either be locally hosted or server based
  - Local databases include:
  - A spreadsheet*
  - A CSV file*
  - A locally hosted database system, such as SQLite, Libré Base or MS Access



```
A:A1: 'EMP                                                                  MENU
Worksheet  Range  Copy  Move  File  Print  Graph  Data  System  Quit
Global  Insert  Delete  Column  Erase  Titles  Window  Status  Page  Hide
A     A         B          C       D         E        F        G
1  EMP      EMP_NAME       DEPTNO  JOB       YEARS    SALARY   BONUS
2      1777 Azibad           4000 Sales          2     40000    10000
3     81964 Brown            6000 Sales          3     45000    10000
4     40370 Burns            6000 Mgr            4     75000    25000
5     50706 Caeser           7000 Mgr            3     65000    25000
6     49692 Curly            3000 Mgr            5     65000    20000
7     34791 Dabarrett        7000 Sales          2     45000    10000
8     84984 Daniels          1000 President      8    150000   100000
9     59937 Dempsey          3000 Sales          3     40000    10000
10    51515 Donovan          3000 Sales          2     30000     5000
11    48338 Fields           4000 Mgr            5     70000    25000
12    91574 Fiklore          1000 Admin          8     35000      ---
13    64596 Fine             5000 Mgr            3     75000    25000
14    13729 Green            1000 Mgr            5     90000    25000
15    55957 Hermann          4000 Sales          4     50000    10000
16    31619 Hodgedon         5000 Sales          2     40000    10000
17     1773 Howard           2000 Mgr            3     80000    25000
18     2165 Hugh             1000 Admin          5     30000      ---
19    23907 Johnson          1000 VP             1    100000    50000
20     7166 Laflare          2000 Sales          2     35000     5000
DATA.WK3
```

Ulster University



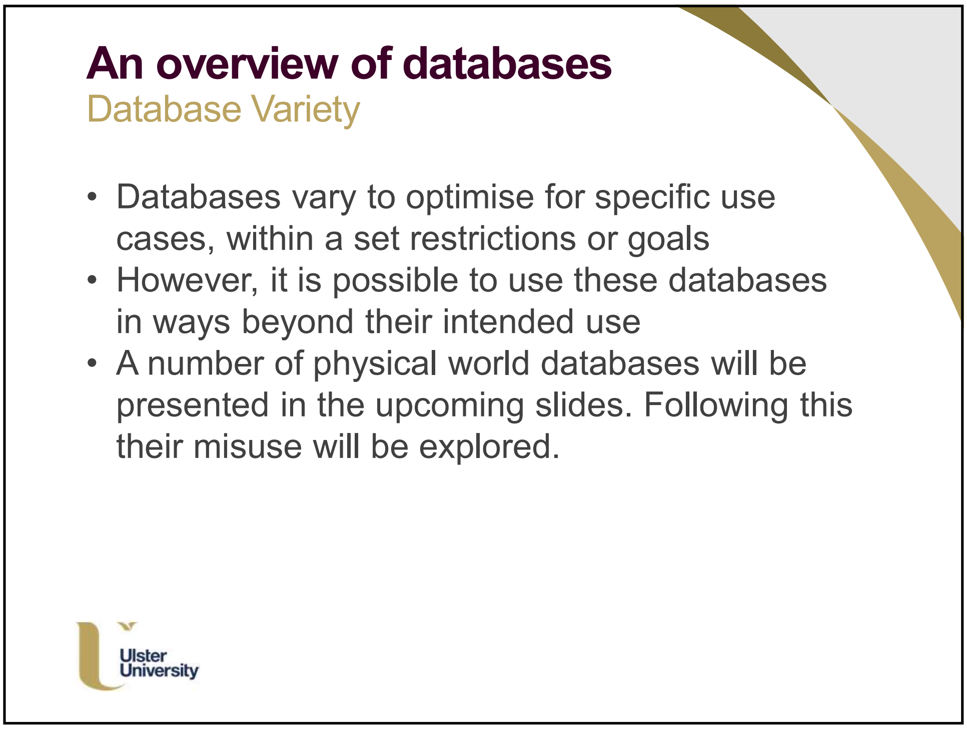
An overview of databases

Introduction

- Server based databases are generally performance optimised, shared, resources
- These have the greatest variety of design goals
- Server based databases include:
  - MySQL – A relational DB
  - InfluxDB – A time series DB
  - MongoDB – A document oriented DB
  - Neo4J – a graph DB

Ulster University

11



An overview of databases

Database Variety

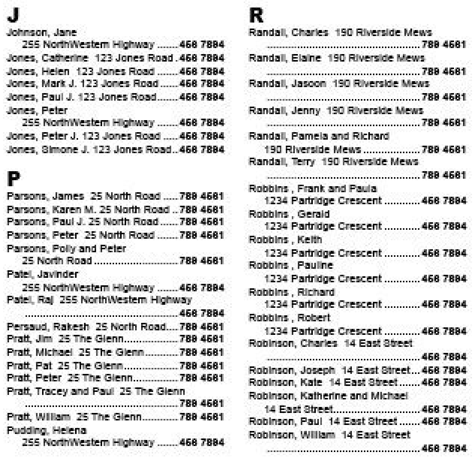
- Databases vary to optimise for specific use cases, within a set restrictions or goals
- However, it is possible to use these databases in ways beyond their intended use
- A number of physical world databases will be presented in the upcoming slides. Following this their misuse will be explored.

Ulster University

12

An overview of databases

Database Variety



J

Johnson, Jane
255 NorthWestern Highway 468 7884
Jones, Catherine 123 Jones Road .468 7884
Jones, Helen 123 Jones Road 468 7884
Jones, Mark J. 123 Jones Road 468 7884
Jones, Paul J. 123 Jones Road.......468 7884
Jones, Peter
255 NorthWestern Highway 468 7884
Jones, Peter J. 123 Jones Road 468 7884
Jones, Simone J. 123 Jones Road..468 7884

P

Parsons, James 25 North Road 788 4681
Parsons, Karen M. 25 North Road ..788 4681
Parsons, Paul J. 25 North Road......788 4681
Parsons, Peter 25 North Road 788 4681
Parsons, Polly and Peter
25 North Road 788 4681
Patel, Javinder
255 NorthWestern Highway 468 7884
Patel, Raj 255 NorthWestern Highway
....468 7884
Persaud, Rakesh 25 North Road....788 4681
Pratt, Jim 25 The Glenn..................788 4681
Pratt, Michael 25 The Glenn...........788 4681
Pratt, Pat 25 The Glenn..................788 4681
Pratt, Peter 25 The Glenn...............788 4681
Pratt, Tracey and Paul 25 The Glenn
....788 4681
Pratt, William 25 The Glenn............788 4681
Pudding, Helena
255 NorthWestern Highway 468 7884

R

Randall, Charles 190 Riverside Mews
....788 4681
Randall, Elaine 190 Riverside Mews
....788 4681
Randall, Jasoon 190 Riverside Mews
....788 4681
Randall, Jenny 190 Riverside Mews
....788 4681
Randall, Pamela and Richard
190 Riverside Mews 788 4681
Randall, Terry 190 Riverside Mews
....788 4681
Robbins , Frank and Paula
1234 Partridge Crescent 468 7884
Robbins , Gerald
1234 Partridge Crescent 468 7884
Robbins , Keith
1234 Partridge Crescent 468 7884
Robbins , Pauline
1234 Partridge Crescent 468 7884
Robbins , Richard
1234 Partridge Crescent 468 7884
Robbins , Robert
1234 Partridge Crescent 468 7884
Robinson, Charles 14 East Street
....468 7884
Robinson, Joseph 14 East Street...468 7884
Robinson, Kate 14 East Street.......468 7884
Robinson, Katherine and Michael
14 East Street.............................468 7884
Robinson, Paul 14 East Street 468 7884
Robinson, William 14 East Street
....468 7884

Ulster University

13

An overview of databases

Database Variety

- These databases have optimisations
- Phonebook
  - Use: retrieving a phone number given an name and contracted address in a combined field.
  - Indexed alphabetically by surname
  - Read only
  - All fields non optional
  - Phone number is the unique key

Ulster University

14

An overview of databases

Database Variety

Ulster University

15

An overview of databases

Database Variety

- These databases have optimisations
- Address book
  - Use: retrieving and storing a in-depth contact information about a person
  - Indexed alphabetically by forename or surname, as dictated by end user
  - Supports insertion of data but not re-indexing/restructuring
  - All fields are optional, field size varies
  - There is no strictly enforced unique key

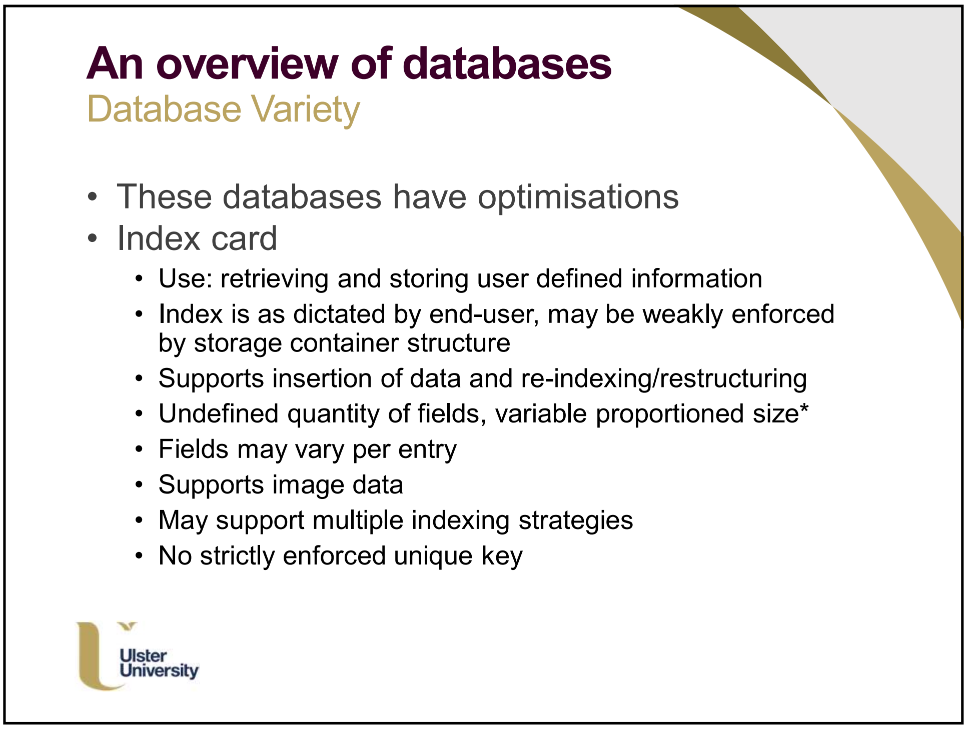
Ulster University

16

An overview of databases

Database Variety

17



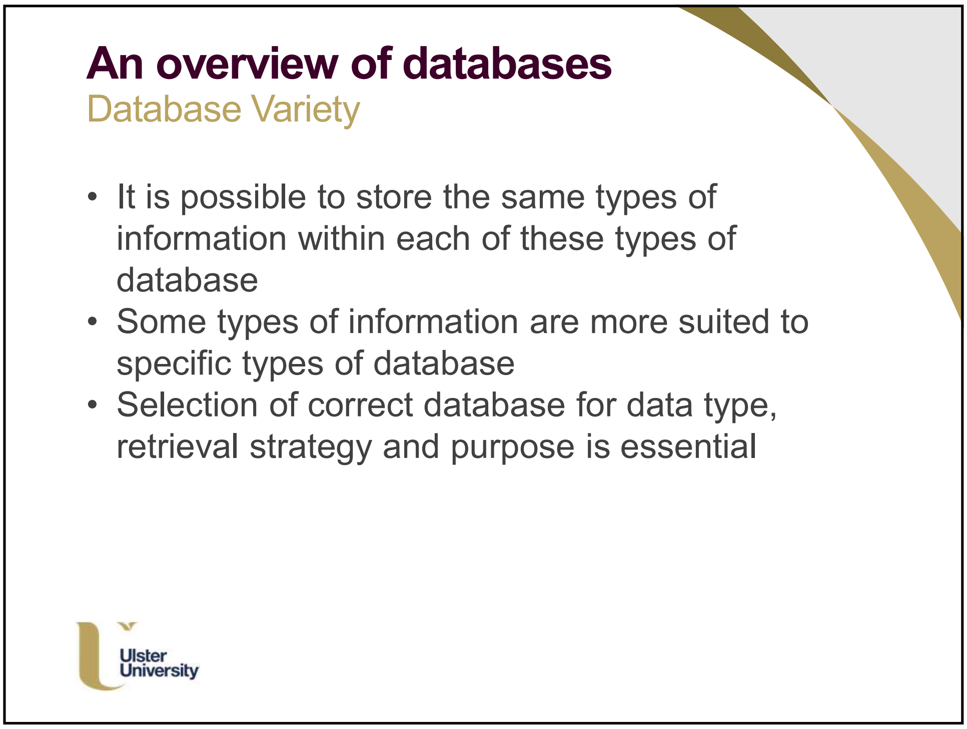
An overview of databases

Database Variety

- These databases have optimisations
- Index card
  - Use: retrieving and storing user defined information
  - Index is as dictated by end-user, may be weakly enforced by storage container structure
  - Supports insertion of data and re-indexing/restructuring
  - Undefined quantity of fields, variable proportioned size*
  - Fields may vary per entry
  - Supports image data
  - May support multiple indexing strategies
  - No strictly enforced unique key

Ulster University

18



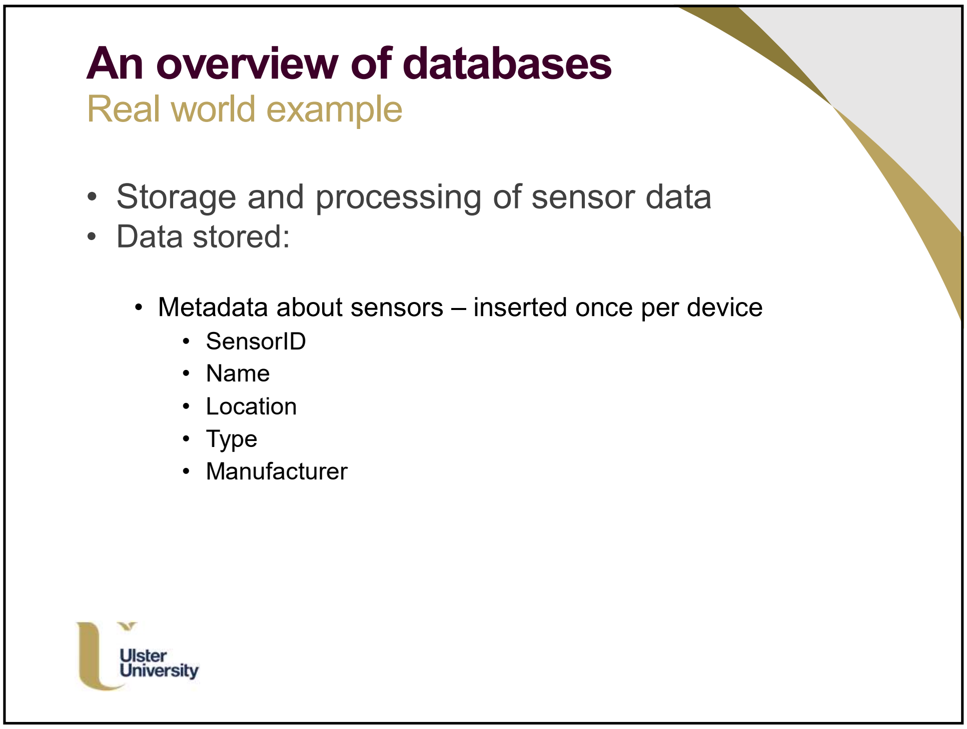
An overview of databases

Database Variety

- It is possible to store the same types of information within each of these types of database
- Some types of information are more suited to specific types of database
- Selection of correct database for data type, retrieval strategy and purpose is essential

Ulster University

19



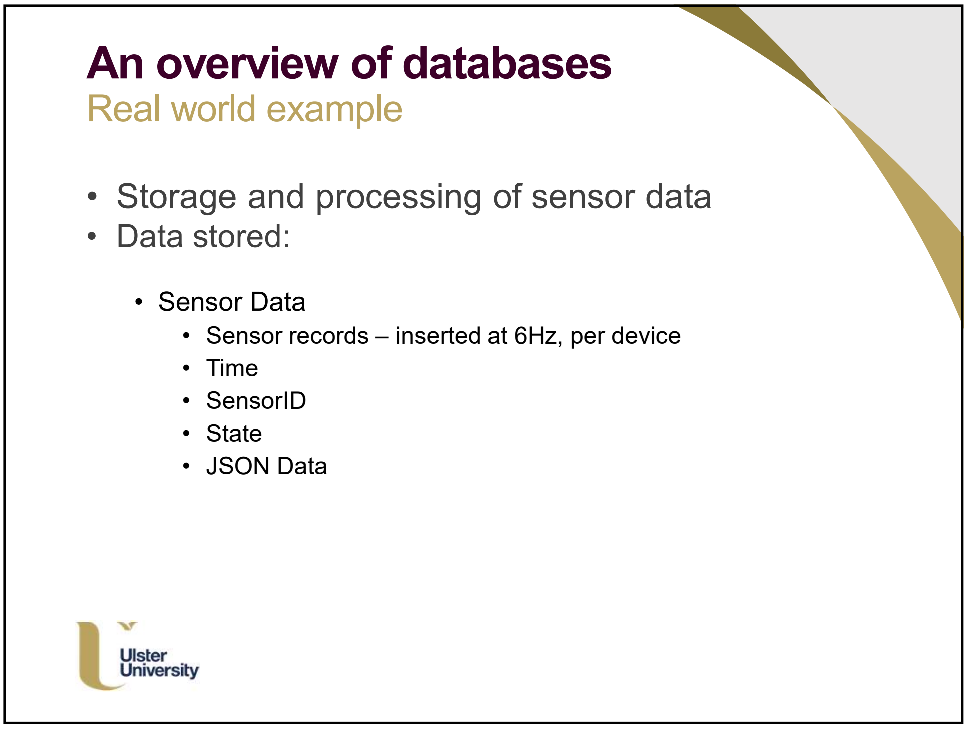
An overview of databases

Real world example

- Storage and processing of sensor data
- Data stored:
  - Metadata about sensors – inserted once per device
    - SensorID
    - Name
    - Location
    - Type
    - Manufacturer

Ulster University

20

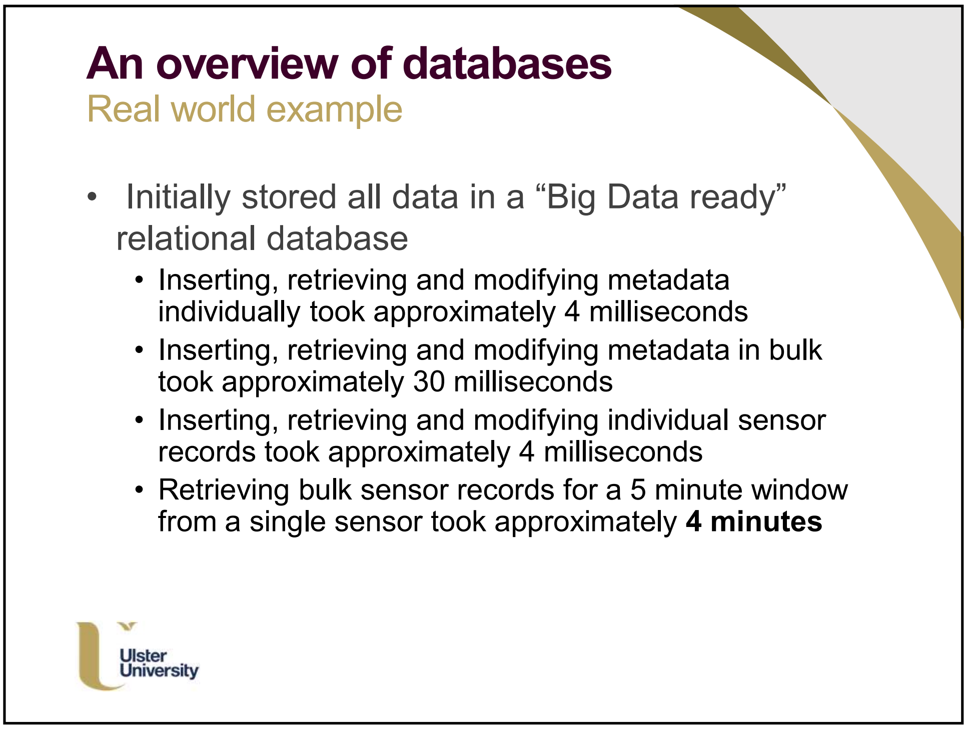


An overview of databases

Real world example

- Storage and processing of sensor data
- Data stored:
  - Sensor Data
    - Sensor records – inserted at 6Hz, per device
    - Time
    - SensorID
    - State
    - JSON Data

21

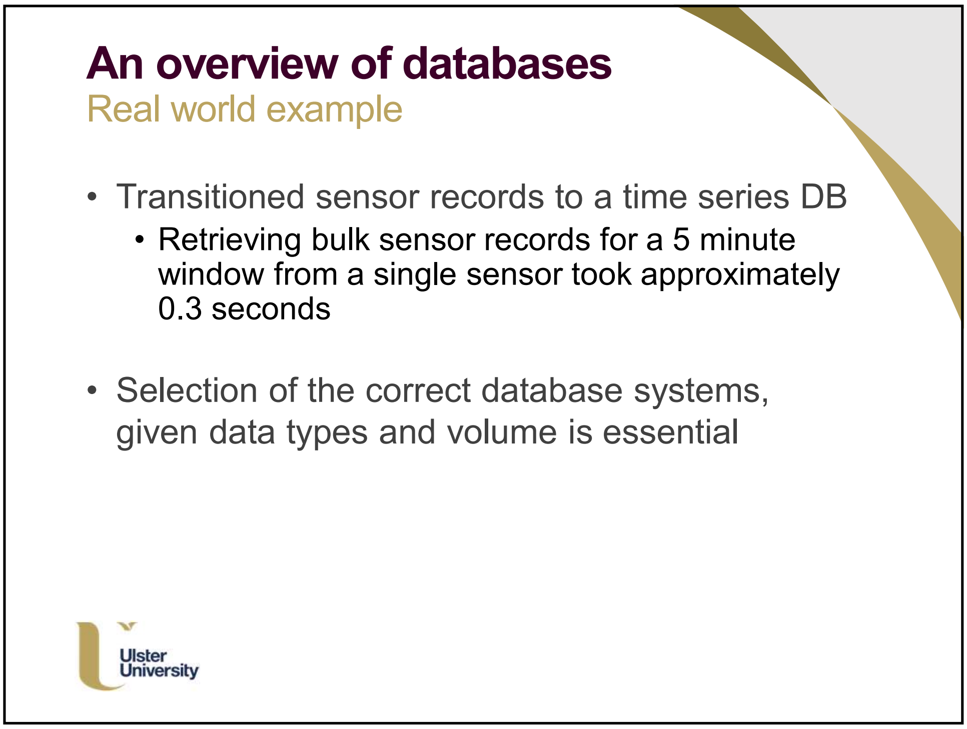


An overview of databases

Real world example

- Initially stored all data in a “Big Data ready” relational database
  - Inserting, retrieving and modifying metadata individually took approximately 4 milliseconds
  - Inserting, retrieving and modifying metadata in bulk took approximately 30 milliseconds
  - Inserting, retrieving and modifying individual sensor records took approximately 4 milliseconds
  - Retrieving bulk sensor records for a 5 minute window from a single sensor took approximately **4 minutes**

22



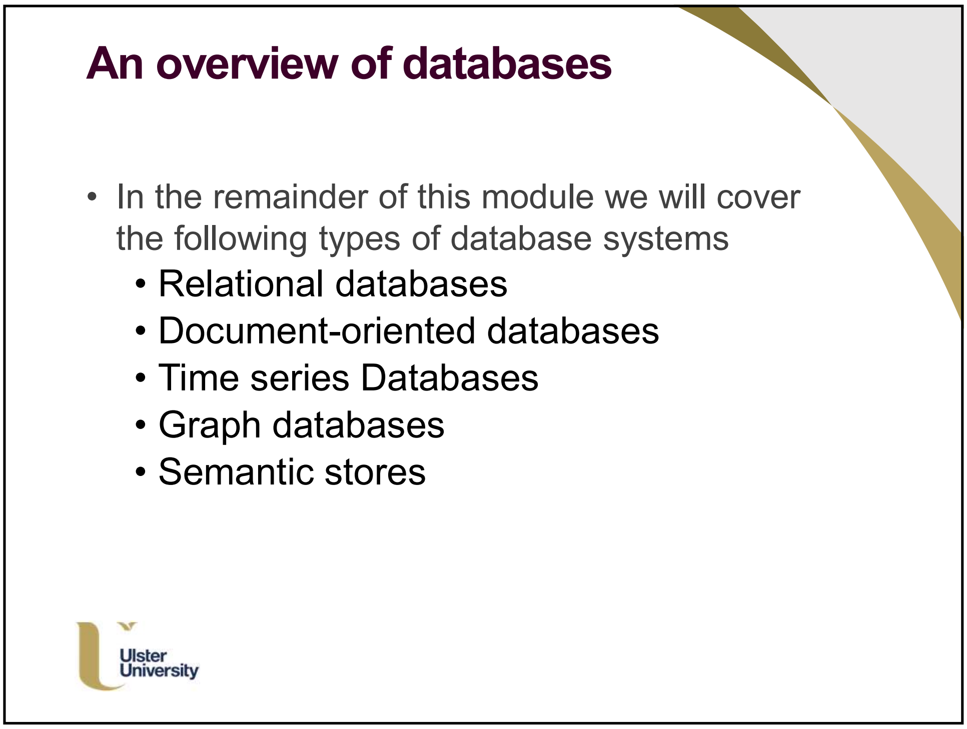
An overview of databases

Real world example

- Transitioned sensor records to a time series DB
  - Retrieving bulk sensor records for a 5 minute window from a single sensor took approximately 0.3 seconds

- Selection of the correct database systems, given data types and volume is essential

23



An overview of databases

- In the remainder of this module we will cover the following types of database systems
  - Relational databases
  - Document-oriented databases
  - Time series Databases
  - Graph databases
  - Semantic stores

24

An overview of databases

- Next topic:

Relational databases

25

Any Questions?

26