Ysgol Reolaeth Gogledd Cymru North Wales Management School

PRIFYSGOL GLYNDŴR WRECSAM WREXHAM GLYNDŴR UNIVERSITY

CONL 722

Big Data Challenges and Opportunities

1.2.1: Database in Big Data Era

This video will explore the database systems and data storage concerning big data.

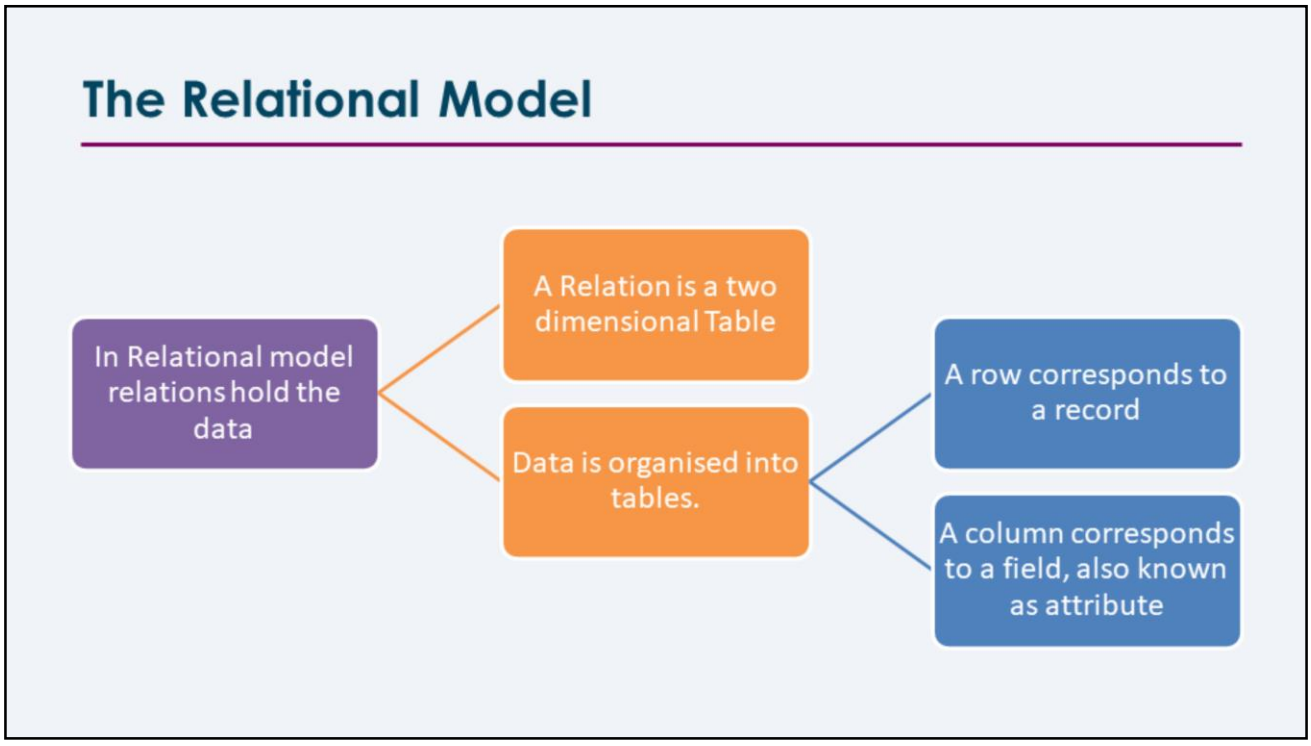
Database

- A Database is an **integrated collection** of stored operational data used by the application systems of a particular enterprise.
  - Data for all applications in the enterprise is stored in the integrated database **not** individual files.
  - **Centralised control** of its operational data

A database can be defined as an integrated collection of stored operational data used by the application systems of a particular enterprise.

This means that the data for all applications in the enterprise is stored in the integrated database NOT in individual files.

This enables the enterprise to provide the CENTRALISED CONTROL for all of its operational data and provide data independence.



The most widely used database system is the relational database system, based on the concept of a Relation,
which is physically represented as a two dimensional table.
Relations will hold information about the 'objects' to be stored in the database.

Data is organised into tables.

A row of the table corresponds to an individual record i.e. the details of an individual object.

The columns in the table correspond to attributes (also known as fields) and represent the characteristics of the object.

The Relational Model

- The relationships tables are represented solely by the data values
- Use JOIN to retrieve the data

COURSE

Course_ID	Course_Name	Level
BSC_COMP	BSc Computing	Undergraduate
MSC_COM_SCI	MSc Computer Science	Postgraduate
MSC_COM_NW	MSc Computer Science with Networking	Postgraduate
MSC_COM_BDA	MSc Computer Science with Big Data Analytics	Postgraduate

STUDENT

Student_ID	First_Name	Last_Name	Course_ID
S001	Jack	Adams	BSC_COMP
S002	Jill	Jones	MSC_COM_SCI
S003	Bob	Who	MSC_COM_SCI
S004	Alice	Davies	MSC_COM_BDA

In the relational model, the relationship between the rows of data in different tables is represented by data values in one or more columns.

Consider the example of you registering for the Online MSc course. (Please note only handful records are used in this examples).
In order to process your registration, your details and course details must be recorded.

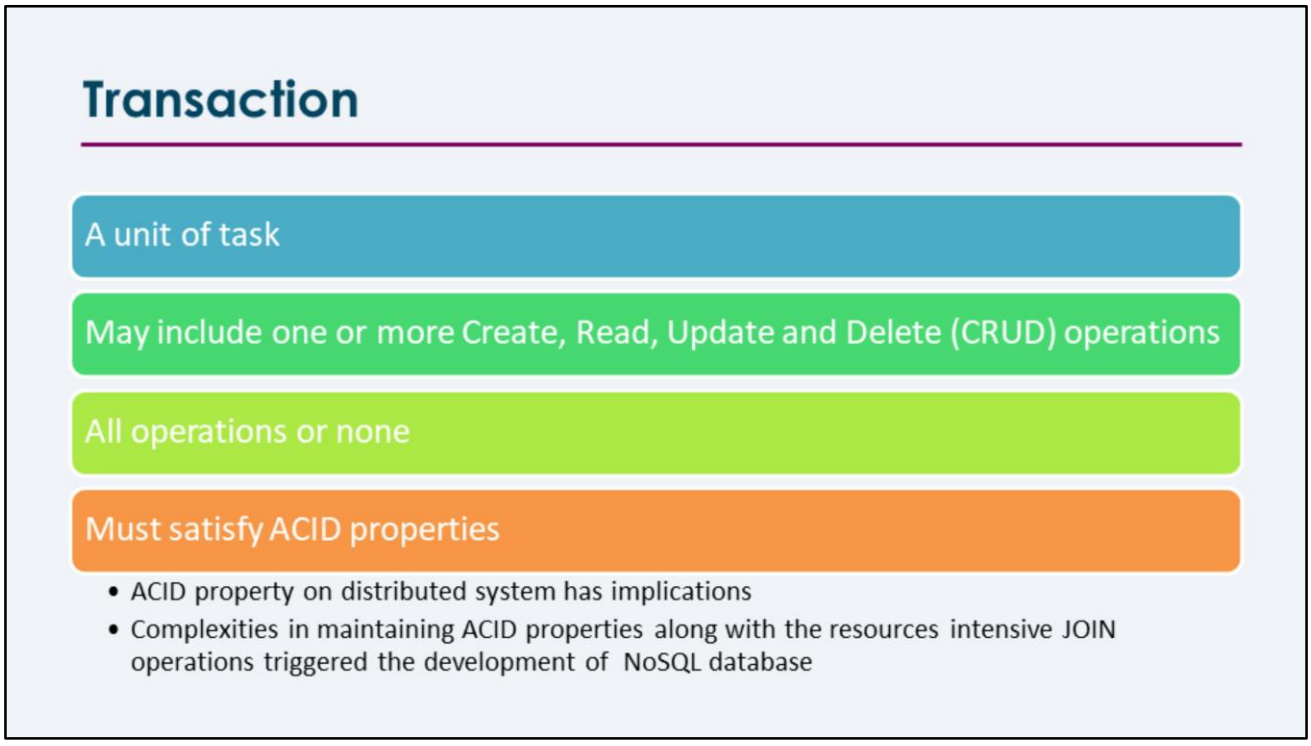
The Student Registration System requires the details of the Course and Details of the Students.

Course and Student will be two objects about which we are collecting the data, hence they are the two tables.

In the database student registering for a course is a relationship which has to be mapped using the data values.

Here you can see the column Course ID from the course table is also appearing in the Student Table.

To retrieve the data from multiple tables, you must JOIN the table using the relationship between the data.



In a database context, an action that is carried out by a user or by an application program is known as a transaction.

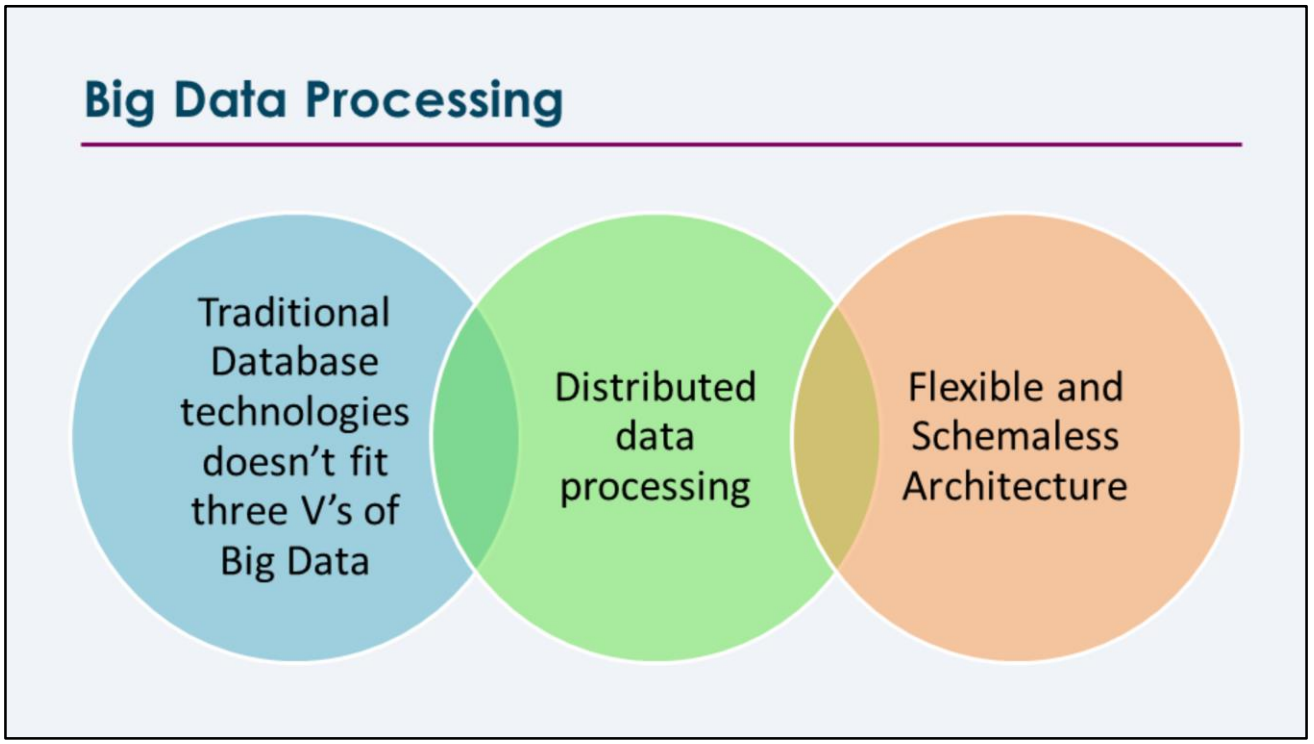
A transaction can be defined as a unit of a task.

It may include one or more Create, Read, Update and Delete (CRUD) operations
Either all of these operations or none of the operations should be completes
Each transaction must satisfy ACID properties as follows:

Atomicity
Consistency

Isolation
Durability

Ensuring ACID property on a distributed system has large implications, it may have to use two-phase commit
Complexities in maintaining ACID properties along with the resources intensive JOIN operations triggered the development of NoSQL database



Traditional centralised database technologies don't fit three V's of Big Data.

Big Data processing required to create a distributed data processing architecture and manage the co-ordination through a programming language.
Distributed data processing can easily handle the large Volume, but not the Variety and Velocity.
In order to handle the 3Vs, we need flexible and schemaless architecture.

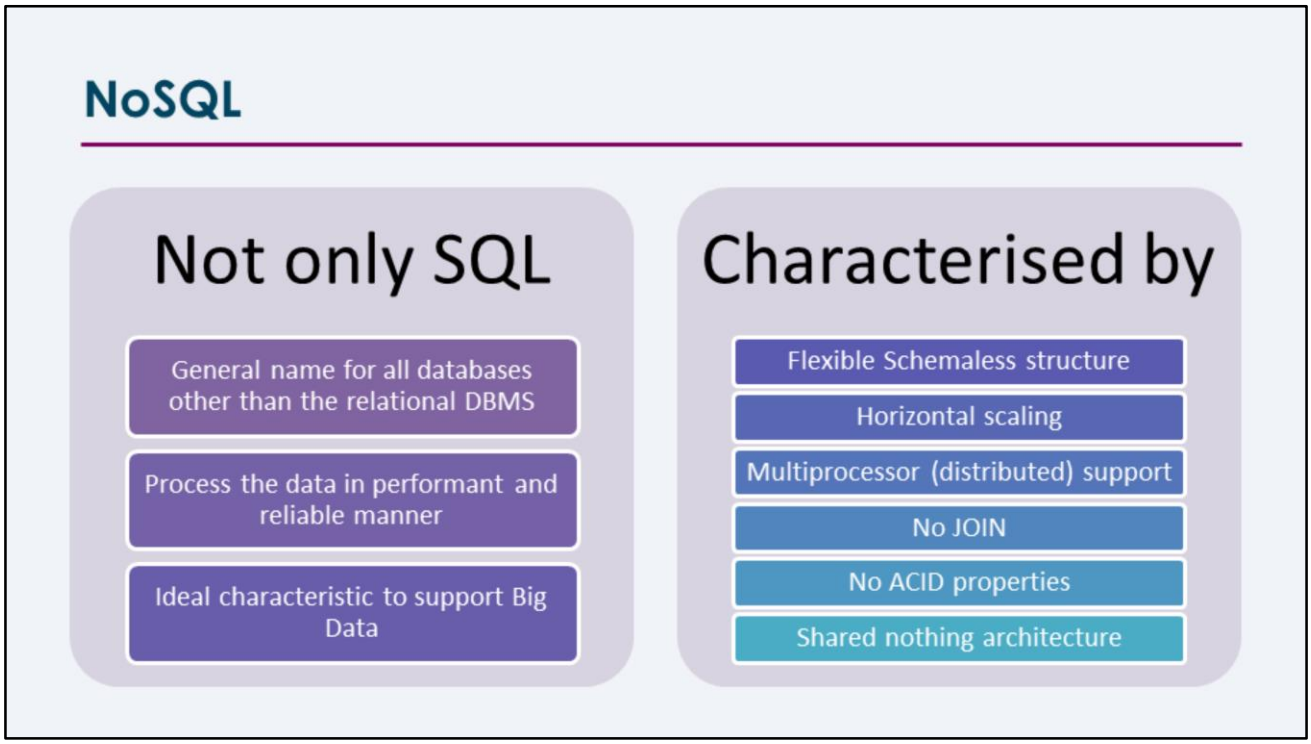
NoSQL

- The proliferation of interactive web application demands storing a large volume of data of varied format
- Relational model relying on predefined normalised structure
  - Capable of handling large quantity
  - Insufficient to support the varied format
- Big Data is characterised by the volume, velocity and variety
  - Relational Database alone is insufficient to support the data storage requirements.

Proliferation of interactive web application demands storing large volume of data of varied format
The relational model relies on the existence of a predefined and normalised structure.

Even though it is capable of handling large quantities of data it is insufficient to support the varied format.

Big Data is characterised by the volume, velocity and variety and relational database alone is insufficient to support the data storage requirements.



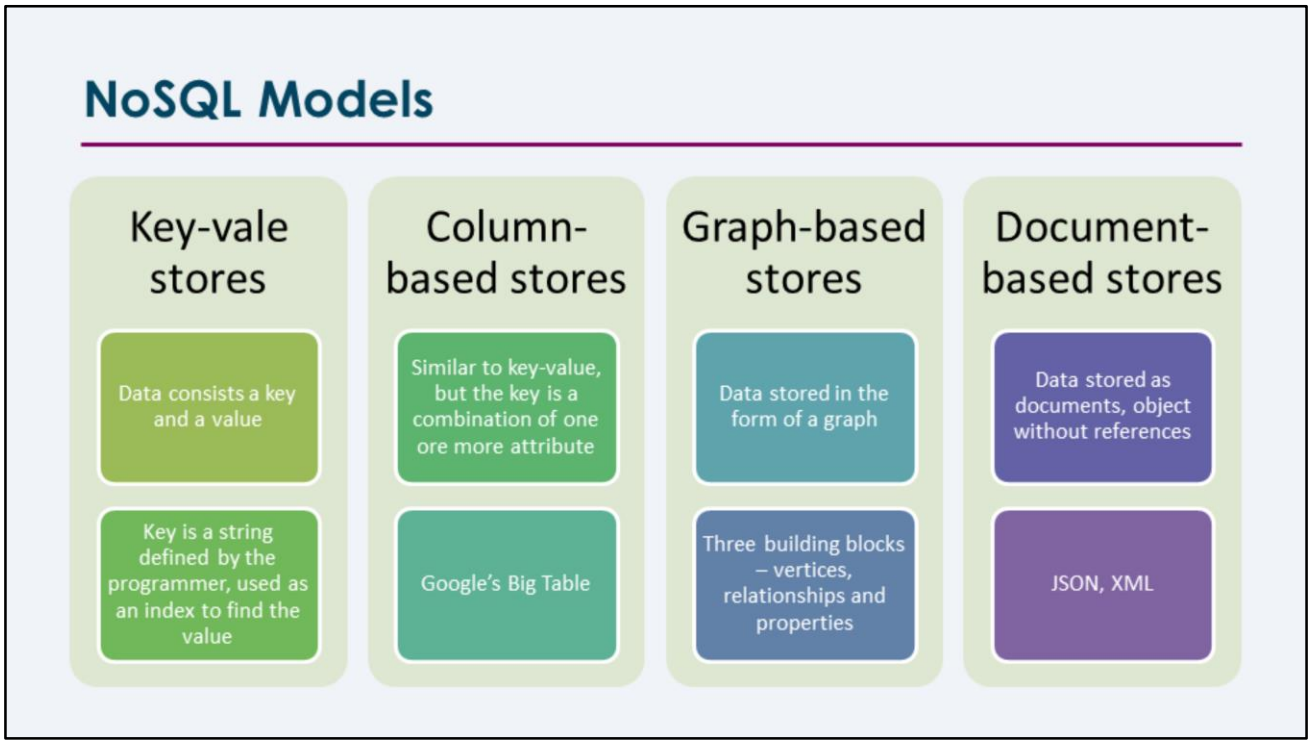
The ever increasing demand for the storage and manipulation of a large volume of data of varying format has supported the NoSQL move.

NoSQL means Not only SQL.

It is the general name for all databases other than the relational database.

It supports processing the data in a performant and reliable manner and it is the ideal characteristic to support Big Data.

NoSQL family of the database is characterised by
Flexible,
Schemaless structure
Horizontal scaling
Multiprocessor (distributed) support
No JOIN operations
Not relying on ACID properties
And support a shared-nothing architecture.



Categorised by the data storage model

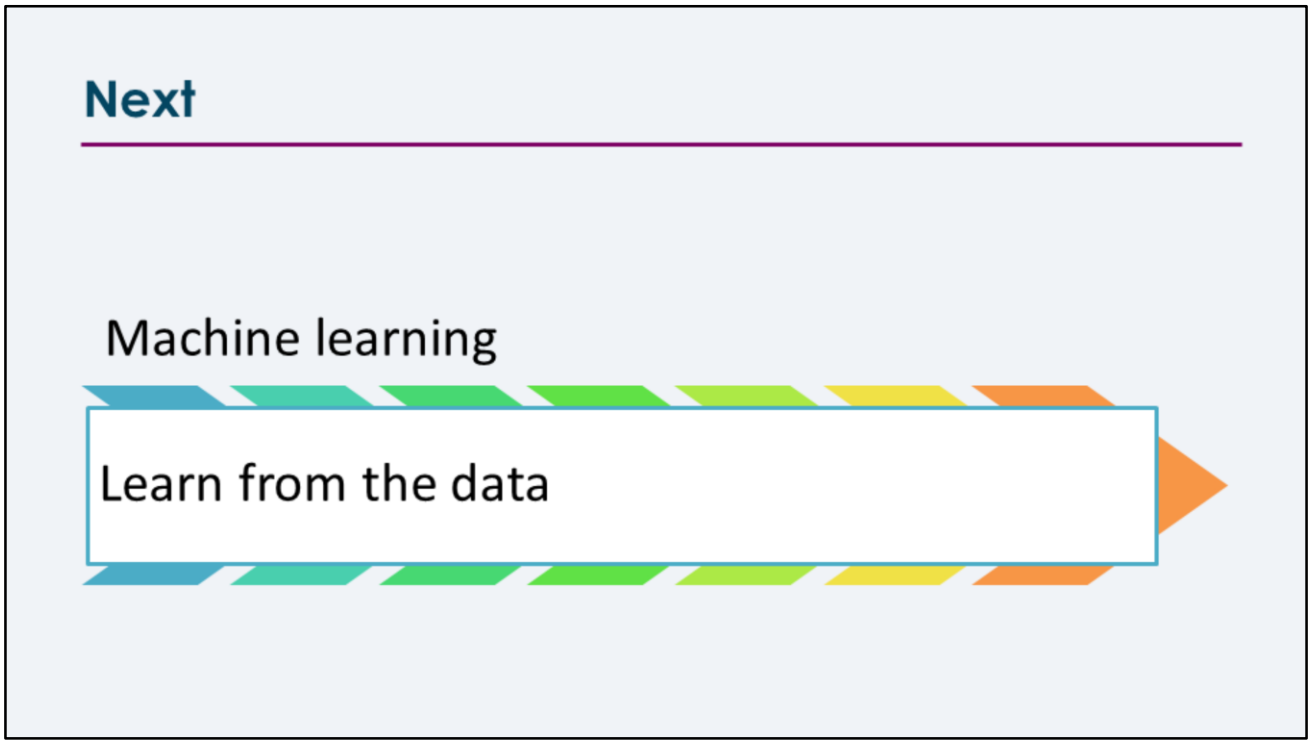
- Key-vale stores
  - Data consists of a key and a value
  - Key is a string defined by the programmer, used as an index to find the value
- Column-based stores
  - Similar to key-value, but the key is a combination of one or more attribute
  - Google's Big Table
- Graph-based stores
  - Data stored in the form of a graph
  - Three building blocks – vertices, relationships and properties
- Document-based stores
  - Data stored as documents, object without references
  - JSON, XML

NoSQL V/s Relational model

- Depends on the application
- Trade off between consistency and availability
  - Some times making the data available is crucial than ensuring consistency
    - Online shopping site want the data available on the webpage about the products rather than ensuring it's the correct stock displayed
      - Discrepancies will be managed by the business
    - Mission critical systems must ensures consistency of the data
  - Can apply CAP theorem
    - Consistency, Availability, Partition tolerance
- Both has its own niche and required by the varied applications and support Big Data analysis in their own way

The selection between NoSQL and Relational model depends on the application.

The trade-off is between consistency and availability.
Sometimes making the data available is crucial than ensuring consistency.
Online shopping site wants the data available on the webpage about the products rather than ensuring it's the correct stock displayed. Discrepancies will be and must be managed by the business.
Mission-critical systems must ensure the consistency of the data.
Can apply CAP theorem
Consistency, Availability, Partition tolerance and evaluate what is the most important factor that should be considered in relation to the application or system in hand.
Both database systems have their own niche and required by the varied applications and support Big Data analysis in their own way.



The next video will be discussing an important component of big data analytics, machine learning. Learning from the data.