Ysgol Reolaeth Gogledd Cymru | North Wales Management School

PRIFYSGOL GLYNDŴR WRECSAM | WREXHAM GLYNDŴR UNIVERSITY

CONL 722

Big Data Challenges and Opportunities

2.2.1: Metadata, Classification and Ontologies

The previous lesson introduced you to structured and unstructured data and highlighted the importance and methods for identifying unstructured data. Unique identifiers will help to distinguish the individual data objects, however, they do not describe the data or represent the association between different data objects. This video will provide an overview of metadata the data descriptor and explain classification and ontologies for grouping related data.

Ysgol Reolaeth Gogledd Cymru
North Wales Management School
PRIFYSGOL GLYNDŴR WRECSAM
WREXHAM GLYNDŴR UNIVERSITY

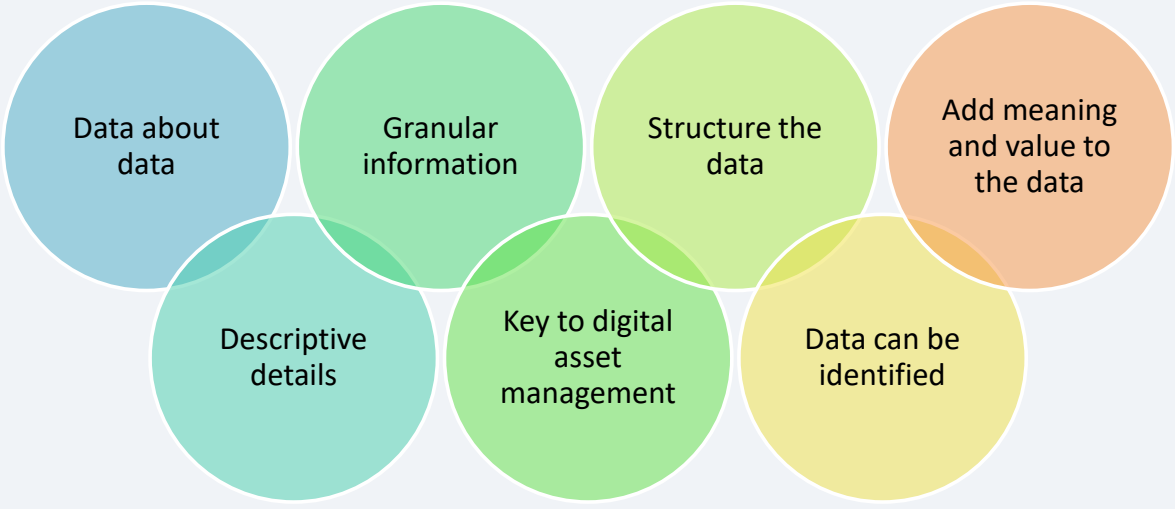
CONL 722

Big Data Challenges and Opportunities

2.2.1: Metadata, Classification and Ontologies

The previous lesson introduced you to structured and unstructured data and highlighted the importance and methods for identifying unstructured data.
Unique identifiers will help to distinguish the individual data objects, however, they do not describe the data or represent the association between different data objects.
This video will provide an overview of metadata the data descriptor and explain classification and ontologies for grouping related data.

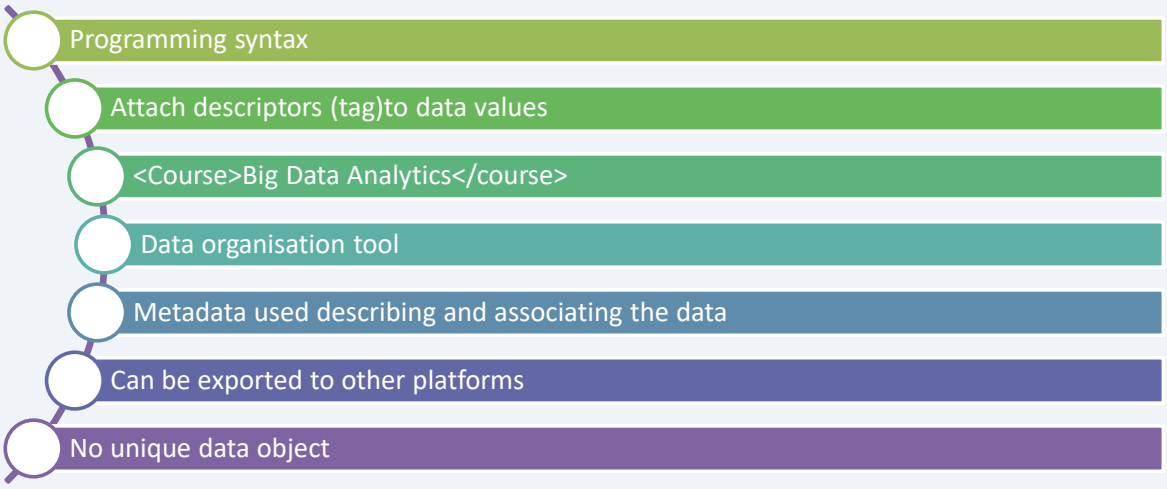
Metadata



Metadata is the data bout the data and provides descriptive details about the data object.
Comparing metadata to big data, we can say metadata can provide information in a granular level where big data can provide an overall trend and pattern.

Metadata is the centre of digital asset management.
It stipulates a structure to the digital asset and provides meaning to the data object.
Appropriate metadata will help to identify the data and understand the meaning of the Data and increase the value of the data. Metadata help to identify the organisational associations of the data.
Comprehensive **metadata** enable the **data** sets designed for a specific purpose to be reused for other purposes.

XML



XML is programming syntax for attaching descriptors to the data.
Descriptors are also known a Tags.
For example, expressing <Course>Big Data Analytics</course> will mean Course is Big Data Analytics
The description 'course' is represented as Tags and the value Big Data analytics is provided in the middle.

XML provides a structure and organise the data into that structure.
Metadata can be used for describing and associating the data as well as for exporting data to other platforms.
Even though it is a powerful data management toll one major drawback is data/metadata pair is not assigned to a unique object.
Using resource description framework data/metadata pair can be associated with a unique data object.

Semantics and Triples

Data descriptors

- <Course>Big Data Analytics</Course>
- <Fee_in_GBP>5000</<Fee_in_GBP>

Individually not very meaningful

- Name
- Fees

What about

- Course Big Data Analytics fee in GBP is 5000?

5000 has meaning with a course name

The previous slides explained metadata, which is the data about data provides structure. XML provides a programming interface to define data using descriptors. However, this explicitly does not provide the relationship between data values. From the example, the course tag identifies Big Data analytics as a course. Fee in GBP tag identifies 5000 is the fees in GBP. Individually they have no meaning. But associating fees to the course as Big Data Analytics' fees in GBP is 5000, have a meaning.

The numeric data 5000 has a meaning when it is associated with a course name.

Semantics and Triples

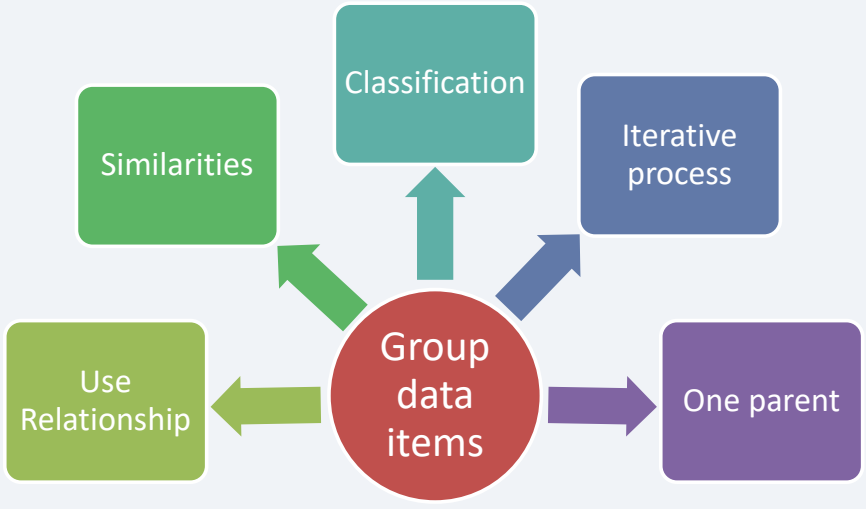
- In Big data associations will be managed using UUID
  - 8a5dgs78-e---------- course Big Data Analytics
  - 8a5dgs78-e---------- Fee 5000
- Study of the meaning
  - Meaningful assertions
  - Data object, descriptor and value
  - Triples

In Big data resources, the associations between data items are managed using UUID.
Course name and fee details become two assertions, both with the same UUID.
Any more information about the course should also have the same UUID. Appropriate processes must be in place to ensure this.

Semantics is the study of meaning, in big data analytics, it is the technique for analysing text and creating meaningful assertions or in other words associating meanings to the data item.
Not every assertion is true, it will be meaningful but may be false.

Semantics can be structured using a 3-item list consisting of the identified data object, a data value and a descriptor of the value and is referred to as Triples.
Similar to the sentences being the fundamental unit of the language, the triple is fundamental to in information systems.

Classification

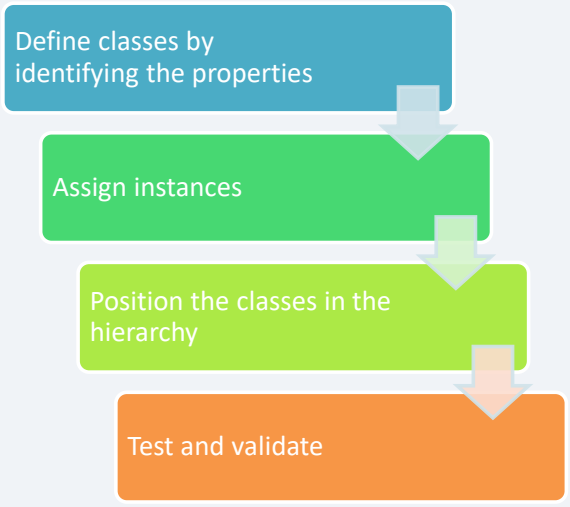


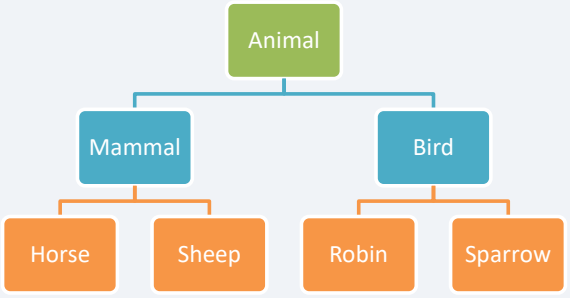
As you have seen in the previous slides, structuring data using databases or identifying data using descriptors alone will not ass any meaning to the data.
Data has meaning when it is associated with other data times and has placed in a context.

As you know big data resources are complex, to harness the power of the data these resources must be divided into classes that share similar properties.
Relationship between the data is one of the fundamental properties used for identifying the class.
Related items may have similarities, but these similarities are the result of relationship rather than a coincidence.
Grouping data items on the relation is classification, where are grouping the data items based on similarity is known as clustering.
Classification is an iterative process, begins with identifying the fundamental properties for a relationship, often generated from hypothesis and verified and revised as the classification grows.
In its simplest form classification will only allow a single parent.

The following slide will briefly explain the steps in building classification.

Classification





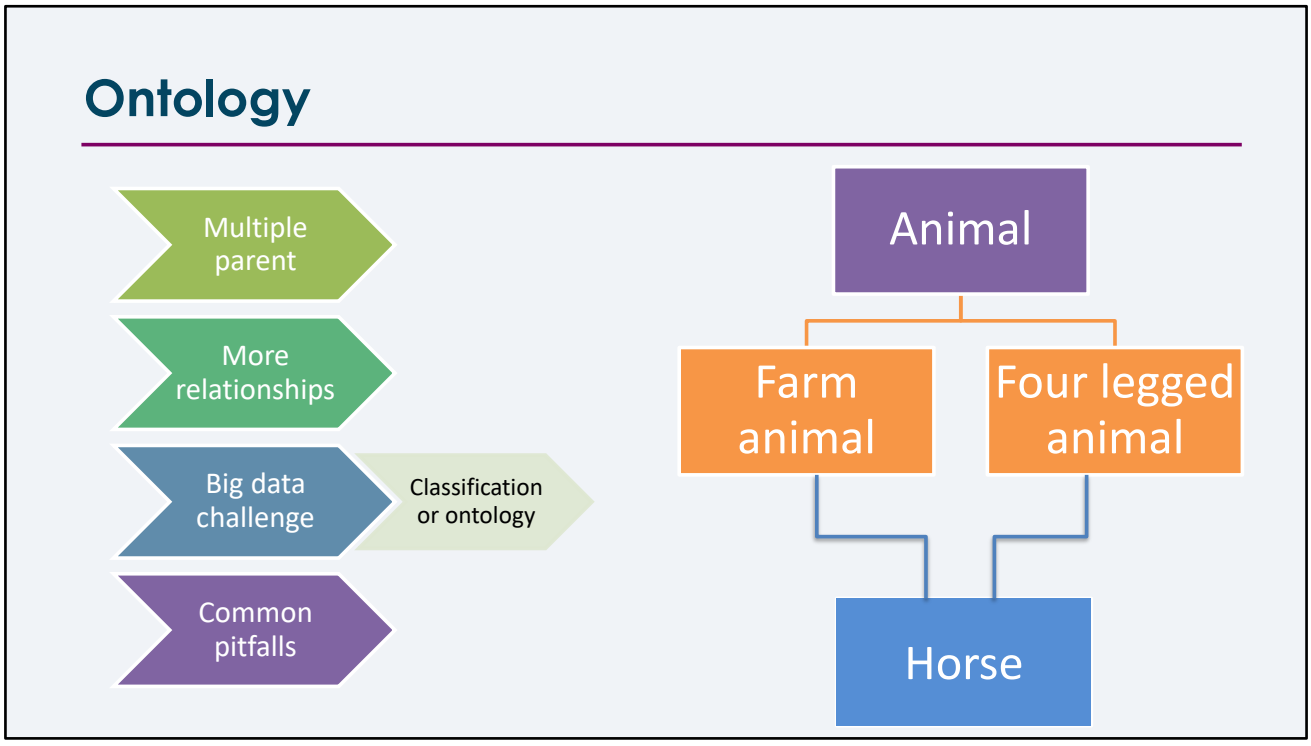
The previous slide stated that the relationship between the data is one of the fundamental properties used for identifying the class.
Hence the first step in classification is to identify the properties of the relationships.

Then identify the subclass, each subclass will only have a single parent (belongs to a single class).
At the top of the hierarchy, the Root will not have any parent.
At the bottom of the hierarchy, the instance will have no more subclass.

A simple example, Animal Root, Mammal, Bird Sub Class, Horse, Sheep, Robin and sparrow are Instances.

Test and validate the classification.

Classification provides a list of every member class along with their relationship to other classes.



Ontologies are different from classification as they allow multiple parents.
An object can be a subclass of more than one class.

Considering the example from the previous slide, animal Class can have subclasses as farm animals, four-legged animals, etc. Horse and sheep could have both parents.
Data analysis prefer ontologies over the classification, as classification limits the possible relationships as it only allows a single parent, whereas Ontologies open up more relationships.

The fundamental difference between classification and an ontology is in the richness of information available. Both provide a list or structure of concepts or classification items.

Classification provides boxes with labels into which to put your data items.

An ontology provides you with a lot more details of concepts, including their relationships. In classification, you place your data in labelled boxes, in ontology, you enrich your data with many relationships stored in the ontology.

Big data resources are complex and have a complex relationship, choosing the right model is a major challenge.

How do you choose classification over ontology? What are the pitfalls? How do you avoid them?
Please refer to the core text, Chapter 5, Classification and Ontologies, explaining these concepts with examples.

Summary

- Structured and Unstructured data
- Identifiers
- De-identifying
- Metadata
- Semantics
- Classification
- Ontologies
- Choosing a model.

During this week you have been exploring some characteristics of the big data resource.
You have realised that the majority of the data sources are unstructured and to process this you have to impose some structure.
The first step in this process is associating the data objects with a Universally Unique Identifier. You have also realised that the data analysis should only be applied to anonymous data, de-identification removes the personalisation of the data when necessary.
Uniquely identified objects can be associated with appropriate metadata giving descriptive details. Descriptors on its own will not provide any meaning to the data, semantics triples will add meaning to data.
Data now can be modelled using classification or ontologies depending on the requirements. Choosing the right model is one of the biggest challenge faced by Big Data managers.

Make sure you read the recommended chapters from the core text.

Now, think about the current COVID situation and think about the role of Big Data analytics focusing on the data collection.

Read the description in the introductory case study looking at the COVID track and Trace system. Try to answer the questions raised in the case study. Engage in the

discussion forum with your responses and commenting on your peers' views.

Next

Big Data Practices

Common approaches and practices in Big Data Application.

Next week will be looking at some common approaches and practices that are followed in a general Big Data analytics platforms.

Looking forward to seeing you all next week. Thank you!