**U6 Lecturecast – Database Normalisation**

Learning outcomes

* Evaluate how the cleaning methods help with the storage of useable datasets;
* Understand how a database is created and how it is linked with the use of key fields;
* Analyse anomalies and how they can affect the integrity of the database;
* Look at normalisation and the reasoning behind the use of different normal forms.

Design one of most critical areas of development.

**Data cleaning**

**Datasets can be better stored in proper models if cleaned. Allows:**

* Easier storage
* Search
* Re-use

Requires attention to detail, understanding of the area of work and clarity on the objective.

Python a tool to clean, format and identify anomalies and issues in the data.

**Standardising and Normalising**

Creating new, more useful values from existing data and adding new data columns.

May also help with identifying outliers and removing – so what is normal range and is there a pattern top the data.

**Primary and Foreign Keys**

**Primary Key** – a unique identifier for an individual record. Could be something that occurs in the record e.g order number, customer ID or it could be a combination (a composite key)

“You can use the primary index to retrieve and access objects from the database. The unique index is a column, or an ordered collection of columns, for which each value identifies a unique row. The sequential values that are assigned in the primary columns determine the order in which the unique index is created”

**IBM, 2021**

**Foreign Key**

A column in a table that references a column in another table (usually primary key) – meant to ensure all references are valid. Ensure references in table match accordingly. For example, join issues reported over time to a single entity.

**Anomalies**

Normalisation aims to prevent anomalies – unexpected errors being introduced into the data.

**Insertion anomalies –** For example creating duplicate details by adding a value a second time e.g same branch, two addresses

**Deletion anomaly –** deleting content that is the only record, unintended loss.

**Modification anomalies –** different versions due to data held in multiple places.

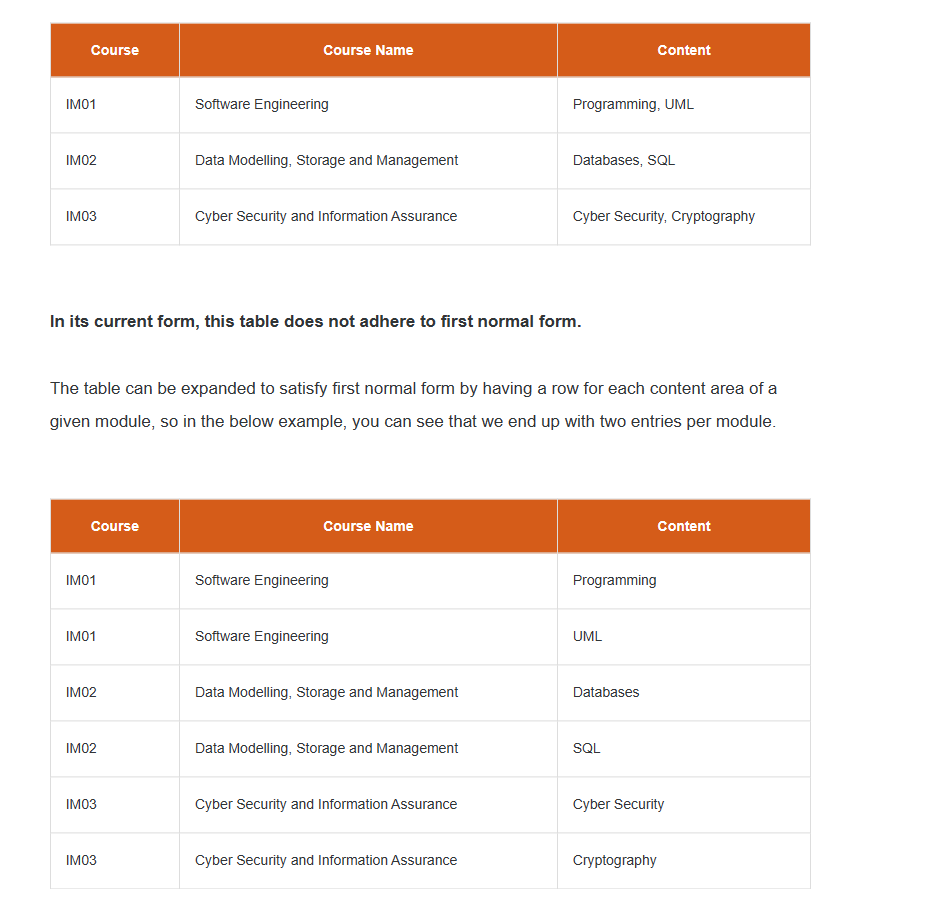
**Normal Forms**

To normalise we are aiming to comply with a number of normal forms.

**First Normal Form**

Achieved if every value atomic and each row unique – each contains only a single value for example, person name split into forename and surname.

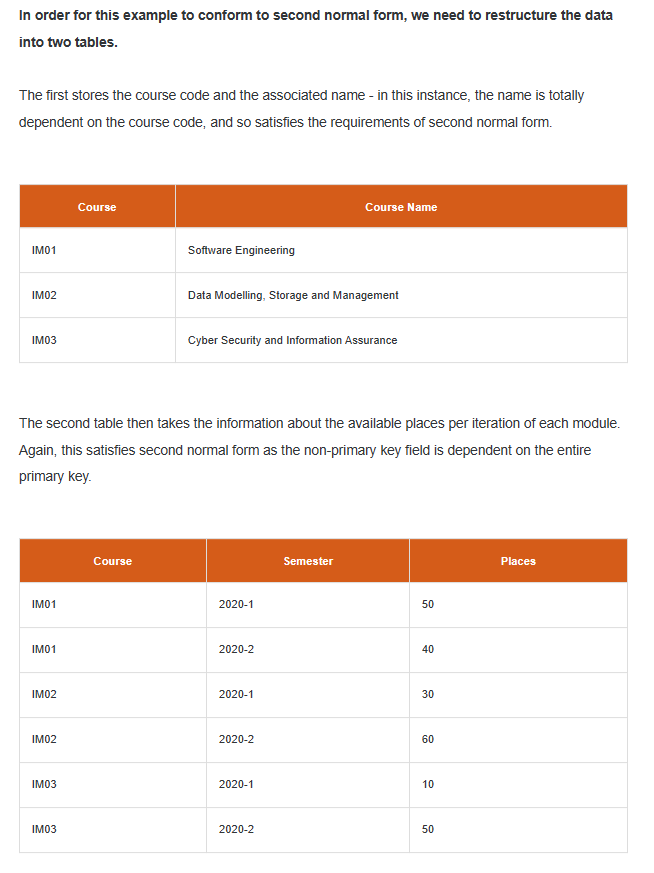
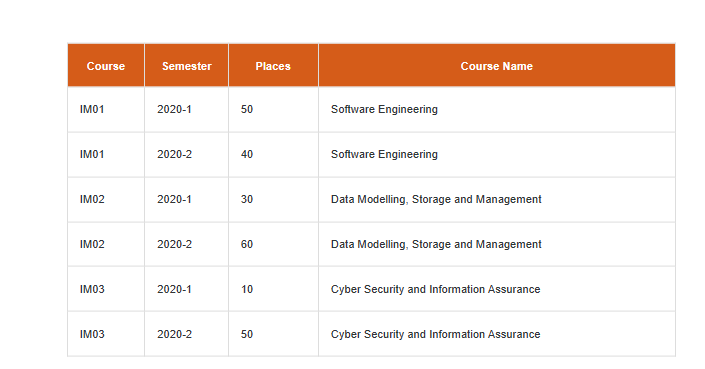
Example from Lecturecast:



**Second Normal Form**

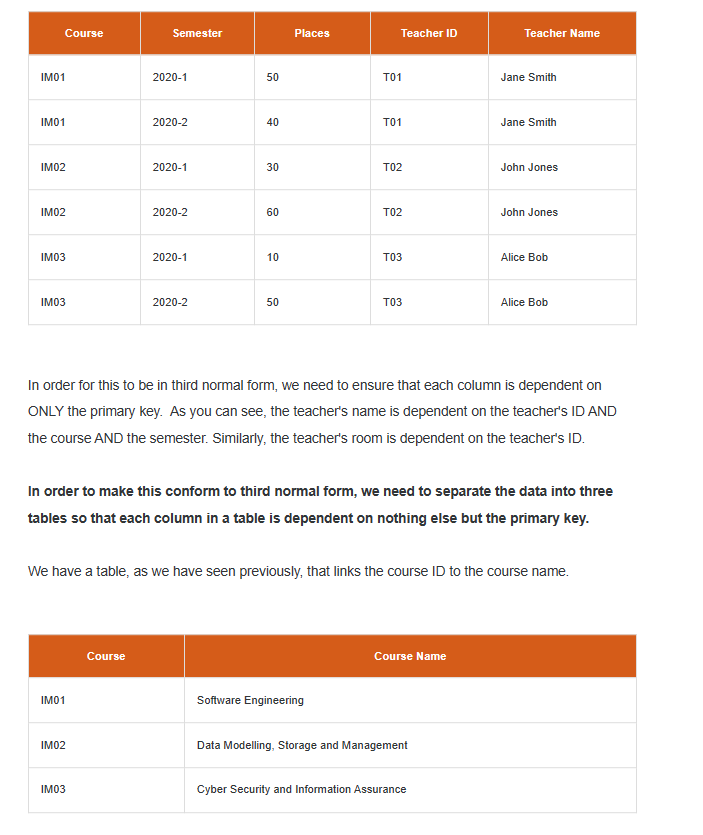
Everything in a table should be dependent and reliant on the primary key (must already be first normal form). May need to split tables into multiple. For example, a data item that is not unique to the entire primary key, for example the name of a data type may only be linked to part of the primary key, then there is specific details such as place linked to the second part. It is not reliant on it only.

Example from lecturecast (primary key is course + semester but the name is only reliant on a part of key)



**Third Normal Formal**Conforms to second normal form and every column not part of the primary key is dependent on the key itself. So effectively split data into a series of tables that provide the detail, almost as a series of lookups between tables:

Example from lecturecast:





"The data depends on the key (1NF), the whole key (2NF) and nothing but the key (3NF)... so help me Codd." Codd, refers to Edgar F. Codd the inventor of relational databases.

**Reflections**

The concepts in this Lecturecast on Normalisation built on my knowledge from working in a professional environment. I have worked with and built data models that apply most of these concepts and can recognise where they have been applied and not applied. The idea of building the database and laying out to avoid the issues with poorly collected data is a strong concept to take from the module as it allows greater access to the data to perform analytics.

**References**

Ganti, V. and Sarma, A. (2013) Data Cleaning : A Practical Perspective (Synthesis Lectures on Data Management). 1 st Ed. Morgan & Claypool Publishers.

Kazil, J. & Jarmul, K. (2016) Data Wrangling with Python. 1st ed. Sebastopol, CA: O’Reilly.

Waldherr, A., Maier, D., Miltner, P., Günther, E. (2017) Big Data, Big Noise. Social Science Computer Review 35(4): 427–443. DOI: 10.1177/0894439316643050

IBM (2021) Primary keys. Available from: https://www.ibm.com/docs/en/controldesk/7.6.0?topic=SSWT9A\_7.6.0/com.ibm.mbs.doc/configur/c\_primary\_key.html