multiverse

multiverse

Data Analysis in Industry

Session 2

Session Outline

Data Types

Data Sources

Data Structures

Data Files

Extract, Transform, Load

Recap

Learning Objectives

- Identify **Business Specific Rules** related to datasets and data characteristics that will influence project design and analysis
- Describe the key characteristics of the different **Data Formats** and how to work with them

multiverse

Data Types

Potential Data Problems

Quantitative

Quantitative

Discrete

Quantitative

Discrete

Continuous

Quantitative

Discrete

Continuous

Qualitative

Quantitative

Discrete

Continuous

Qualitative

Binomial

Quantitative

Discrete

Continuous

Qualitative

Binomial

Nominal

Quantitative

Discrete

Continuous

Qualitative

Binomial

Nominal

Ordinal

Quantitative

Discrete

Continuous

Numerical data that can be 'counted'

e.g. number of marbles, siblings, customers, etc

Quantitative

Discrete

Continuous

Numerical data that can be 'measured'

e.g. temperature, weight, height

Categorical data that has two options

e.g. true or false, heads or tails, yes or no

Qualitative

Binomial

Nominal

Ordinal

Qualitative

Binomial

Nominal

Ordinal

Categorical data that has multiple options but no implied order

e.g. colour, job title, error type, etc

Categorical data that multiple options and an implied order

e.g. likert scale, coffee cup size, salary band, etc

Qualitative

Binomial

Nominal

Ordinal

Identify the Qualitative Data

Weight of a baby	Emotional state	Colour of a bottled drink
Political opinion	Your height	Number of shoes you own
Car type	Holiday destination	Distance to your nearest shop
Number of classes on a timetable	Movie rating	IQ score

Identify the Qualitative Data

Weight of a baby	Emotional state	Colour of a bottled drink
Political opinion	Your height	Number of shoes you own
Car type	Holiday destination	Distance to your nearest shop
Number of classes on a timetable	Movie rating	IQ score

Activity

In groups discuss data you use regularly and whether it is quantitative or qualitative

- What subdivision does it fall under?
- How do you visualise it?
- How do you use it?

Data Sources

What is a data source?

A data source is the location where data is extracted from

Public Data

Client Data

Proprietary Data

Research Data

Public Data

Open Data

Data that can be moved freely, reused and redistributed, although hard to change or modify

Public Data

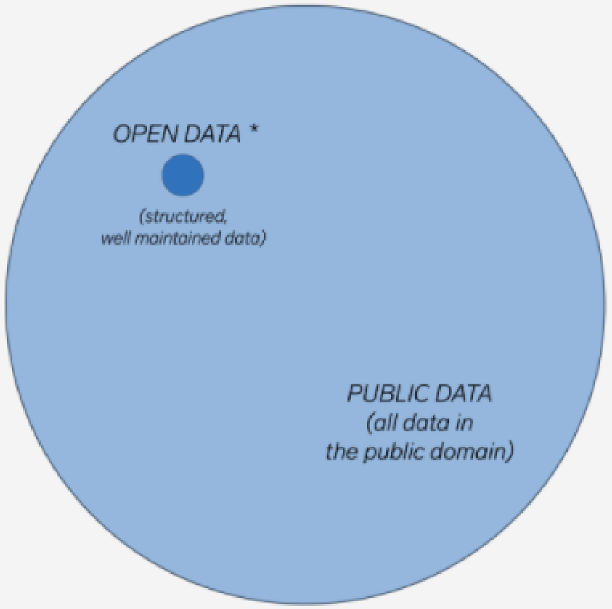
Open Data

A subset of public data but:

- Smaller in volume
- More likely to be structured
- More likely to be open licensed
- Better maintained and more reliable through sanctioned portals
- May require a nominal fee to be used

According to the Open Knowledge Foundation:

“Open data and content can be freely used, modified, and shared by anyone and for any purpose.”


OPEN DATA *
(structured, well maintained data)
PUBLIC DATA (all data in the public domain)

* According to the Open Data Barometer's Global Report 2017, only **7%** of key datasets across 115 countries were considered open. The open data circle size is **7%** of data otherwise considered public.

Proprietary

Operational

Administrative

Data that is owned and stored within an organisation. Proprietary data may be protected by patents, copyrights/trademarks or trade laws.

Proprietary

Operational

Administrative

Proprietary data that is produced by your organisations day to day operations.

E.g. customer, inventory or purchase data

Proprietary

Operational

Administrative

Required to run an organisations day to day operations

E.g. HR, payroll, admin

Client

Proprietary data provided by a client

E.g. data provided by a consultancy firm

Research

Observational

Simulation

Derived

Data from a third party that is made available to you under a licence agreement or has been collected, generated or created to validate original research findings.

Research

Observational

Simulation

Derived

Data gathered from observing trends in the population or from experiments

For example, are shoppers more likely to buy items at eye level?

Research

Observational

Simulation

Derived

Data gathered from a theoretical experiment based on past information

For example, simulating what will happen to the housing market if interest rates rise.

Research

Observational

Simulation

Derived

Data that has been created from other sources

For example, a data warehouse created with ETL

Things to consider:

Things to consider:

Data Accuracy - Can we trust this data? Is it up to date? Is it relevant?

Things to consider:

Data Accuracy - Can we trust this data? Is it up to date? Is it relevant?

Limitations of Data - Are things excluded?

Things to consider:

Data Accuracy - Can we trust this data? Is it up to date? Is it relevant?

Limitations of Data - Are things excluded?

Compatibility with other data sources - Can we join this to our data?

Things to consider:

Data Accuracy - Can we trust this data? Is it up to date? Is it relevant?

Limitations of Data - Are things excluded?

Compatibility with other data sources - Can we join this to our data?

Legal & regulatory rights to data - Are we allowed to use this data?

Things to consider:

Data Accuracy - Can we trust this data? Is it up to date? Is it relevant?

Limitations of Data - Are things excluded?

Compatibility with other data sources - Can we join this to our data?

Legal & regulatory rights to data - Are we allowed to use this data?

Business Context - Do we understand the quirks of this data?

Data Protection Act (DPA 1998)

1	Data must be kept secure
2	Data stored must be relevant
3	Data stored must be kept no longer than necessary
4	Data stored must be kept accurate and up to date
5	Data must be obtained and processed lawfully
6	Data must be processed within the data subject rights
7	Data must be obtained and specified for lawful purposes
8	Data must not be transferred to countries without adequate data protection laws

General Data Protection Regulation (GDPR 2018)

1	Data must be processed **lawfully, fairly and transparently**
2	Data must be collected for **specified, explicit and legitimate purposes**
3	Data must be **adequate, relevant and limited to what is necessary** for processing
4	Data must be **accurate and kept up to date**
5	Data must be **kept only for as long as is necessary** for processing
6	Data must be processed in a manner that **ensures its security**

Activity

Compare the DPA and GDPR

- What is similar?
- What is different?
- How does your organisation ensure compliance?
- How does it affect your role?

In Summary

For data protection you should consider:

In Summary

For data protection you should consider:

You have a lawful basis for processing

In Summary

For data protection you should consider:

You have a lawful basis for processing

You are being transparent

In Summary

For data protection you should consider:

You have a lawful basis for processing

You are being transparent

You are processing it properly

When things go wrong...

Facebook Data Breach
July 2017-September 2018

29 Million people affected

British Airways Hack
August 2018 - September 2018

380,000 people affected

Data Structures

	STRUCTURED DATA	UNSTRUCTURED DATA
CHARACTERISTICS:	▪ Pre-defined data models ▪ Usually text only ▪ Easy to search	▪ No pred-defined data model ▪ May be text, images, audio, video or other formats ▪ Difficult to search
STORED IN:	▪ Relational databases ▪ Data warehouses	▪ Applications ▪ NoSQL databases ▪ Data lakes
GENERATED BY:	▪ Humans or machines	▪ Humans or machines

	STRUCTURED DATA	UNSTRUCTURED DATA
APPLICATION EXAMPLES:	▪ Online reservation system ▪ Inventory control ▪ CRM systems ▪ ERP systems	▪ Word processing ▪ Presentation software ▪ Email clients ▪ Media editing tools
DATA EXAMPLES:	▪ Dates ▪ Product names and numbers ▪ Customer name ▪ Error code ▪ Transaction information	▪ Text files ▪ Audio files ▪ Video files ▪ Images ▪ Emails and reports

Structured

Unstructured

Highly organised

Easily read by machines

YEAR	SITES	PARTICIPATION	MEALS SERVED
1968	0.9	56	0.2
1969	1.2	99	0.3
1970	1.9	227	1.8
1971	3.2	569	8.2
1972	6.5	1080	21.9
1973	11.2	1437	26.6
1974	10.6	1403	33.6
1975	12.0	1785	50.3
1976	16.0	2453	73.4

YEAR	SITES	PARTICIPATION	MEALS SERVED
1968	0.9	56	0.2
1969	1.2	99	0.3
1970	1.9	227	1.8
1971	3.2	569	8.2
1972	6.5	1080	21.9
1973	11.2	1437	26.6
1974	10.6	1403	33.6
1975	12.0	1785	50.3
1976	16.0	2453	73.4

YEAR	SITES	PARTICIPATION	MEALS SERVED
1968	0.9	56	0.2
1969	1.2	99	0.3
1970	1.9	227	1.8
1971	3.2	569	8.2
1972	6.5	1080	21.9
1973	11.2	1437	26.6
1974	10.6	1403	33.6
1975	12.0	1785	50.3
1976	16.0	2453	73.4

Structured

Unstructured

Cannot be processed using conventional tools

Be careful!

Sometimes data looks structured but isn't. For example, Excel spreadsheets have no rules around usage, so you can have multiple tables or different data types in one column.

STRUCTURE	FEATURES
File	▪ Used to store information ▪ Used by computers to read and write information that needs to be processed ▪ Organised into record
List	▪ Contains elements of different data types ▪ E.g. ('John', 10, 7.2, True)
Array	▪ Data can be identified by their index position ▪ Similar to a list but can have multiple dimensions ▪ A 2 dimensional array is a matrix
Table	▪ Typical data files with labelled columns (fields) and rows (records)
Tree	▪ Hierarchical collection of data with parent and child nodes

Activity

Discuss whether the data you use regularly is structured on unstructured

Further reading:

Data Lake vs Data Warehouse
SQL vs NoSQL
Structured vs Unstructured Data

Data Files

Activity

Open each of the files and discuss what the defining features are of each

- What do you think the benefits are?
- What about limitations?
- Do you think they are easier for a human or computer to read?
- Which tools/software can you use with each?

Extensible Markup Language
Comma Separated Values
Text File
Rich Text Format
Excel
JavaScript Object Notation

FILE FORMAT	PROPERTIES	BENEFITS	LIMITATIONS
.xml (e**X**tensible **M**arkup **L**anguage)	A hierarchy based markup language that uses user defined keywords to tag data	▪ Easily read by machines ▪ Portable to many different systems	▪ Hard for humans to read ▪ Large size due to repeated markups
.csv (**C**omma **S**eparated **V**alues)	Tabular data separated by commas. Is a raw text value	▪ Lightweight ▪ Easily read by many applications	▪ If there are commas within the data they need to be ‘text qualified’ so interpreter knows they are not delimiters
.rtf (**R**ich **T**ext **F**ormat)	A file that is stored as Raw text but has a markup language to denote basic formatting such as bold, underline etc.	▪ Fairly lightweight ▪ Suitable for holding documents, not actual data	▪ Rarely used ▪ Hard to read due to markups ▪ Used only for wordpad

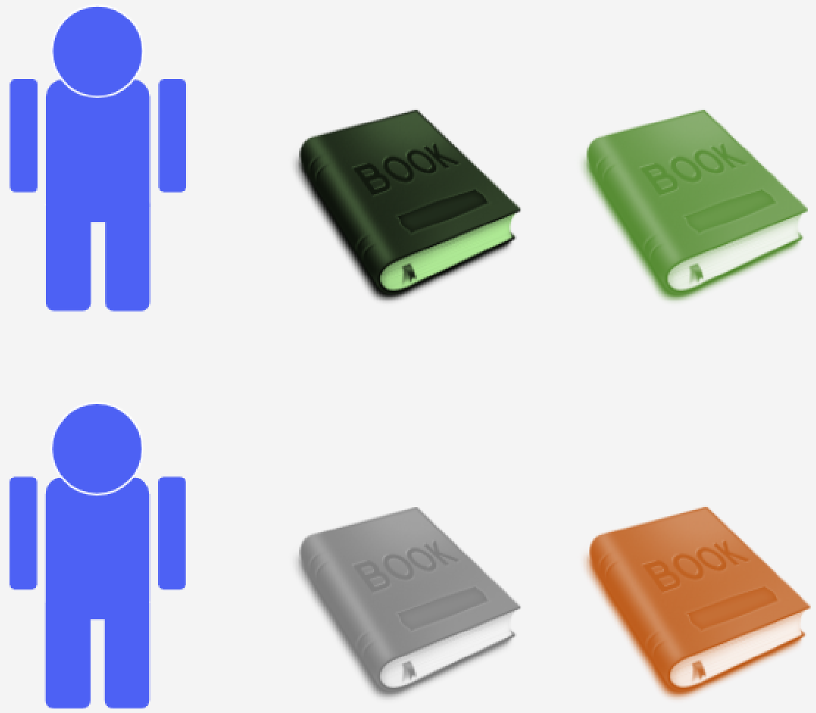
FILE FORMAT	PROPERTIES	BENEFITS	LIMITATIONS
.txt (Text)	Text-based with no formatting or tags. Can be delimited by anything.	▪ Flexible ▪ Lightweight ▪ Easily read	▪ Can easily break ▪ Needs text qualification
.xlsx (Excel File)	Proprietary spreadsheet file format created by Microsoft Excel	▪ Many users are comfortable with this format ▪ Widely used	▪ Large file size ▪ Specialist software needed to view or edit ▪ Hard for applications to read
.json (**J**ava**S**cript **O**bject **N**otation	Text-based open standard designed for human-readable data interchange.	▪ Structure easily read by applications ▪ Lightweight	▪ No error handling ▪ Can leave your machine vulnerable to attacks if taken from an untrusted source

Extract, Transform, Load

How would you **count the number of occurrences** of each word in all the books found in a library **using a team of people?**

Step 1

Divide the books among the team so every person has an allocation



Step 2

Each person will keep a record of the occurrences of each word in their allocation

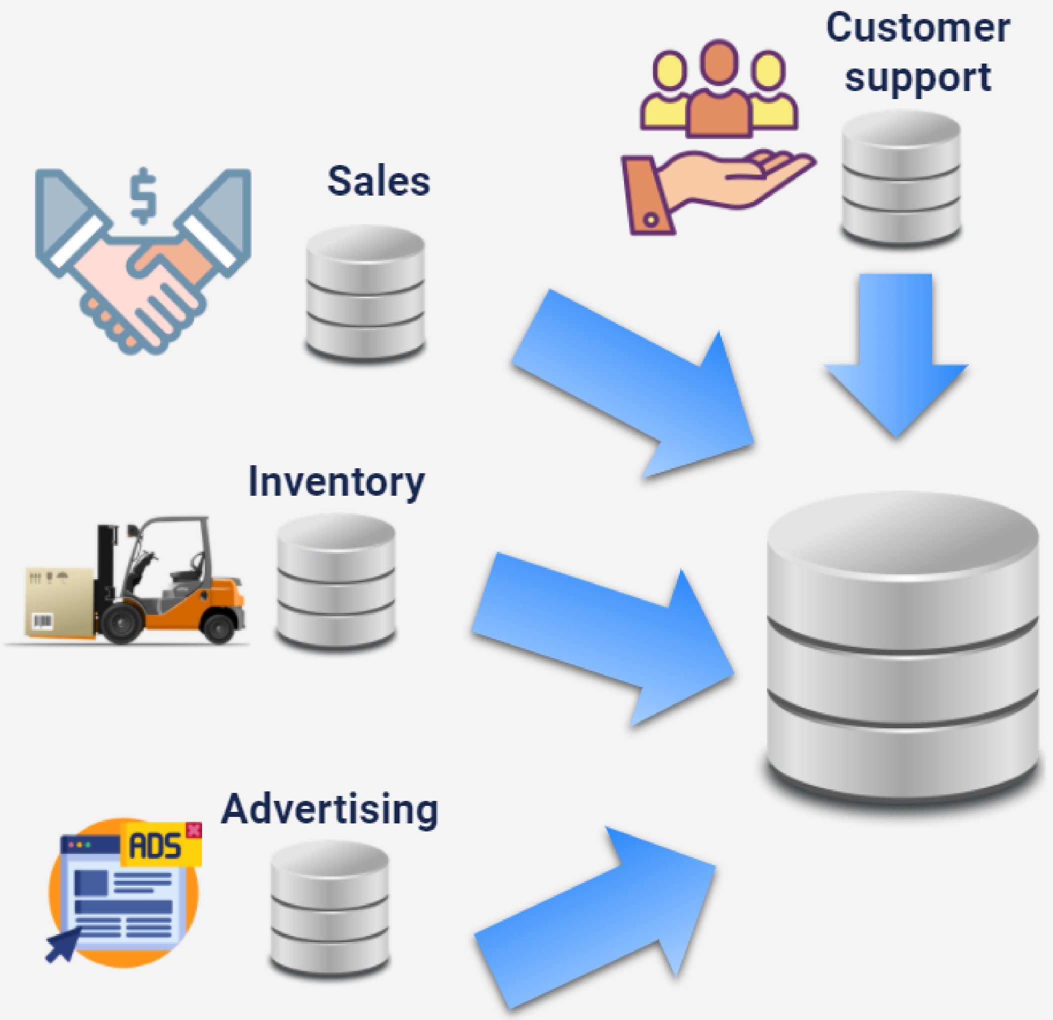
WORD	COUNT
Apple	2
Bird	7

WORD	COUNT
Apple	5
Bird	1

Step 3

Finally combine the different records into one unified view which contains each word in the library.

WORD	COUNT
Apple	7
Bird	8



Customer support

Sales

Inventory

Advertising

Extraction is the process of gathering data from a variety of disparate sources

The extracted data is usually copied from the source, not moved

Validation occurs at this stage to ensure the data is in the correct structure and format, as well as ensuring necessary permissions have been given

The process can be continuous or done in batches

Extract

Transform

Load

Transformation is the process of ensuring the extracted data is in a consistent format

This can include removing null values, changing data types and ensuring field names are the same

As the extracted data is a copy, the original will remian unchanged

Extract

Transform

Load

Extract
Transform
Load

Loading is the process of joining the transformed data together into a single unified view (called the target)

Data verification is undertaken post loading to ensure the combined data is accurate and fulfils the necessary business requirements

With 'Big Data' this process is done using parallel processing to manage the large volume of data being written to the system

Benefits

Allows for a unified view of data that is otherwise spread out across an organisation

Ensures data consistency across an organisation allowing for missing data and errors to be identified throughout a pipeline

Encourages collaboration across teams

Better business intelligence and insights for making decisions through greater data availability

Information Structure and Rules

Data integration activities for data warehouses requires that you follow some basic rules:

- Security policies must be specified by organisations providing data sources to **prevent data leakage and unauthorised access**
- **Access layers** (e.g. networks, firewalls, servers, etc) between sources and targets should be properly configured (especially of data is sourced externally)
- Integrated data should be **immutable**- you should not be able to change the data once it is stored in the unified view
- **Validation checks** should be carried out during ETL:
  - Source and target table structures and data types should be **consistent**
  - **Column names** should be the same as defined by a mapping document

Information Structure and Rules

Data integration activities for data warehouses requires that you follow some basic rules:

- **Verification** is also carried out on the target:
  - Verify that the data is **accurate**
  - Verify the data is the **'right' data** to be stored in the target
  - Verify the data has **not been duplicated**

Recap

Learning Objectives

- Identify **business specific rules** related to datasets and data characteristics that will influence project design and analysis
- Describe the key characteristics of the different **Data Formats** and how to work with them

ASSIGNMENT

PART 1- DATA ANALYTICS LIFE CYCLE

Use a work-related example to identify the stages of the Data Analytics Lifecycle. Describe what happened in each stage and highlight what was your role in the process. In the end, add a summary of the project/analysis including the main findings, what went well and what could have been improved.

Word Count	Max 1500 words
Deadline	3 weeks
Deliverables	Word Document or PowerPoint presentation

ASSIGNMENT

PART 2- PROJECT BRIEF

Use a work-related example to create a project brief. This could be related to a project you are about to start or something new. Your brief should contain a business problem, the wider context of the analysis and a plan of action to solve the problem.

Word Count	Max 1500 words
Deadline	4 weeks
Deliverables	Word Document

Complete Session Attendance Log and Update Your OTJ