IDT403: Lecture Notes

Data management in manufacturing

Unit objectives

- **Understand** the **value of data** in the manufacturing supply chain.
- **Understand** and **remember** the concept of big data and the complexity of managing it.
- **Understand** the **concepts** of relational and non-relational (NoSQL) databases and that different types of data require different storage solutions.
- **Understand** the **concept** of **data modelling** and how to use it.
- **Understand** the **difference** between a data warehouse and a data lake and **when to use** each.
- **Understand** the **concepts** of data preparation, tidying and transforming data.

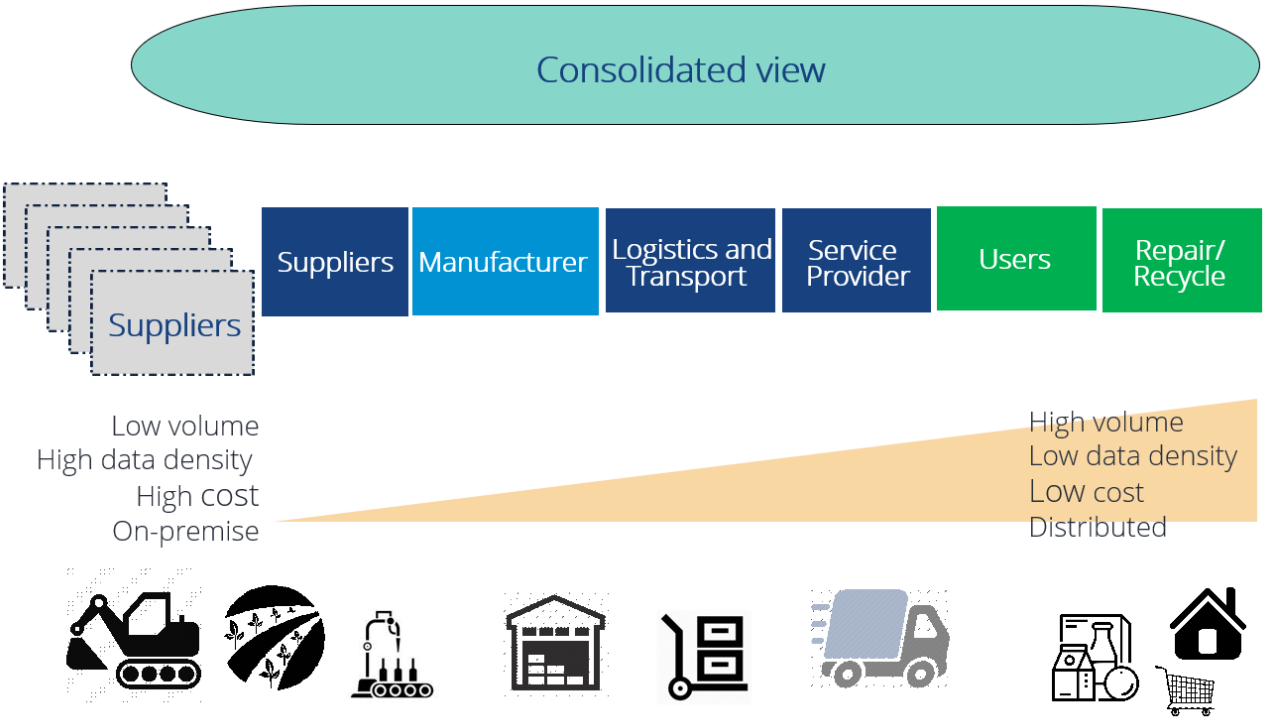
1) Data used in manufacturing

Welcome to the second unit of the introduction to data analysis and machine learning course. In the first section you will explore data used in manufacturing. You will be introduced to different types of data used in the manufacturing context and how data is typically stored and processed. You will also explore data management and data preparation and we will have a discussion on data related aspects regarding your use cases previously selected.

The term “data” typically refers to information in the form of observations related to a particular phenomenon. In the context of computer science and analytics the term however often refers to how information is stored on computers. It can take the form of text, images, audio, or other. Data derive from many sources in the manufacturing context e.g., sensor measurements, events, text, images, and videos. Manufacturing sensor data, otherwise known as the Internet of Things (IoT) generates streams of information. Much of this data is unstructured: images are a collection of pixels with each pixel containing RGB (red, green, blue) colour information. Texts are sequences of words and nonword characters, often organized by sections, subsections, and soon. Clickstreams are sequences of actions by a user interacting with an app or web page.

OGL National Composite Centre 2022
This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

Industry 4.0 requires consolidated data across the supply chain as value can be derived from collecting, storing, transforming, integrating, and analysing data across manufacturing supply chains.

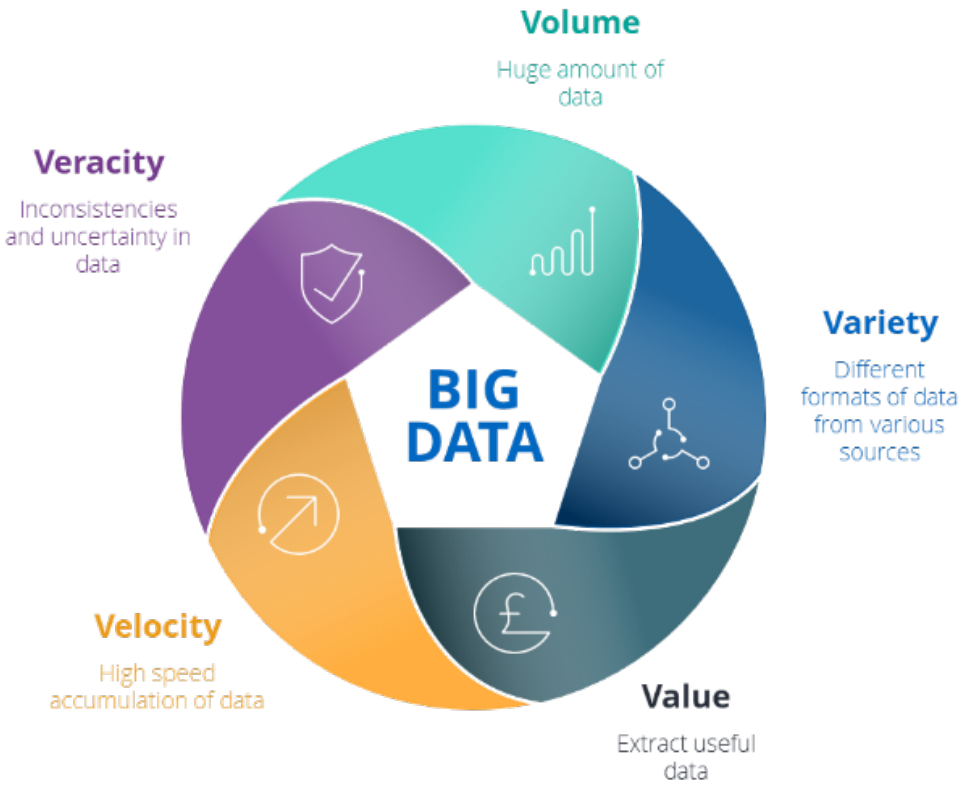


Some of the key challenges to achieving this consolidation are:

- The inability to extract knowledge from the data (value)
- The unstructured format of collected data (variety)
- The massive data to manage, store, and process (volume)
- Insufficient quality of the collected data (veracity)
- Insufficient data processing power (velocity)

These challenges are typical to environments where large amounts of data are generated and this particular application of data analytics have been named “Big Data” as a consequence.

OGL National Composite Centre 2022
This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

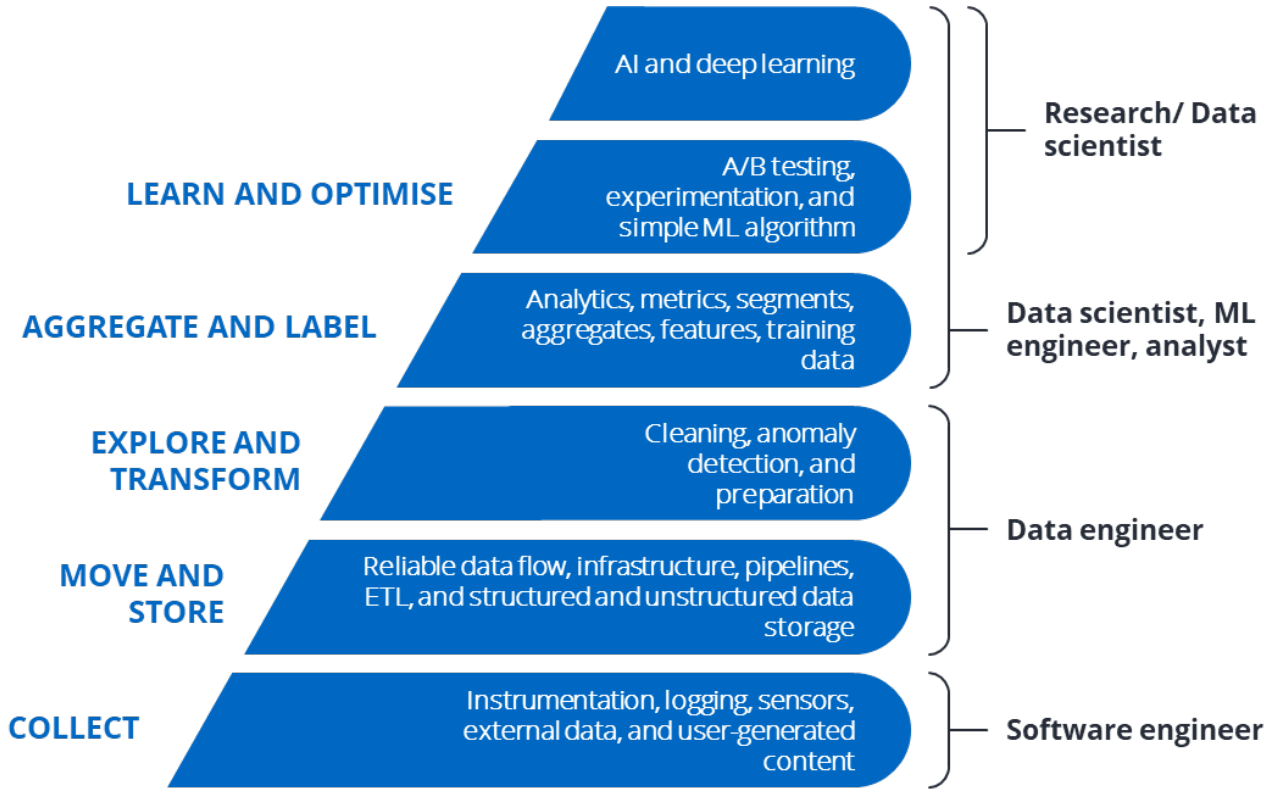


Another problem with manufacturing data is that it is siloed. Not only are different parts of the manufacturing supply chain siloed, but there are also siloes within supply chain units/ organisations, some of the reasons are as follows (Bajic et al., 2021):

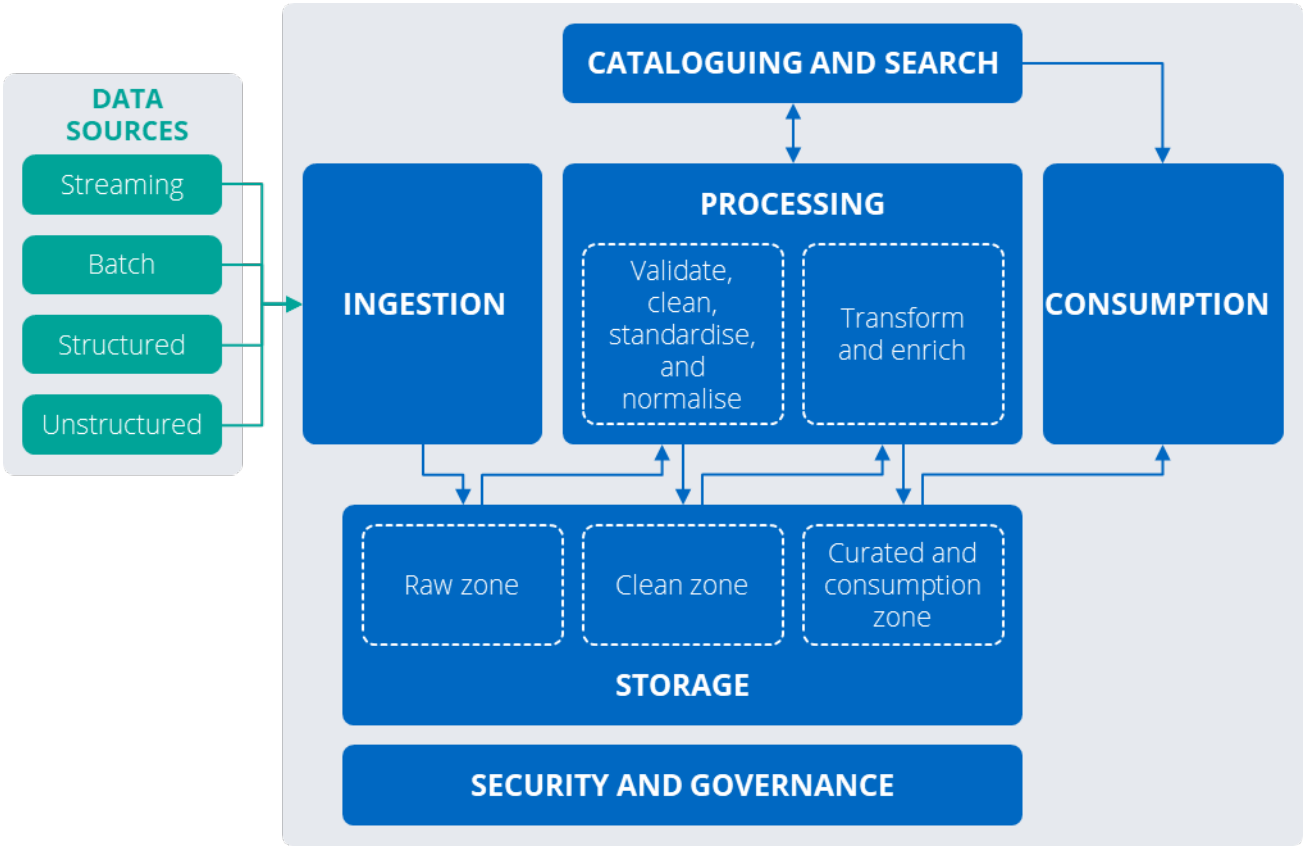
- Many manufacturing systems/ equipment run on proprietary software with proprietary data management solutions, making data integration difficult.
- Legislation challenges in sharing data across organisation boundaries/ within an organisation.
- Lack of information policies and governance to support the process of information sharing.
- Lack of skills to integrate data at various levels of data value pyramid.
- Organisational challenges to adopt integrated solutions/ change management.
- Uncertainty around costs of large-scale data storage.

To address these issues, we need to recognise the different categories of tasks that need to be performed in a typical data science workflow. A typical data science workflow is: “collect, store, explore, transform, analyse and apply”. The following diagram depicts these tasks as well the relevant areas within manufacturing that apply, and the skills required to address the tasks.

OGL National Composite Centre 2022
This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0



A possible cloud-based data management framework that can assist with these tasks is depicted below.



The architecture supports various types of data as input:

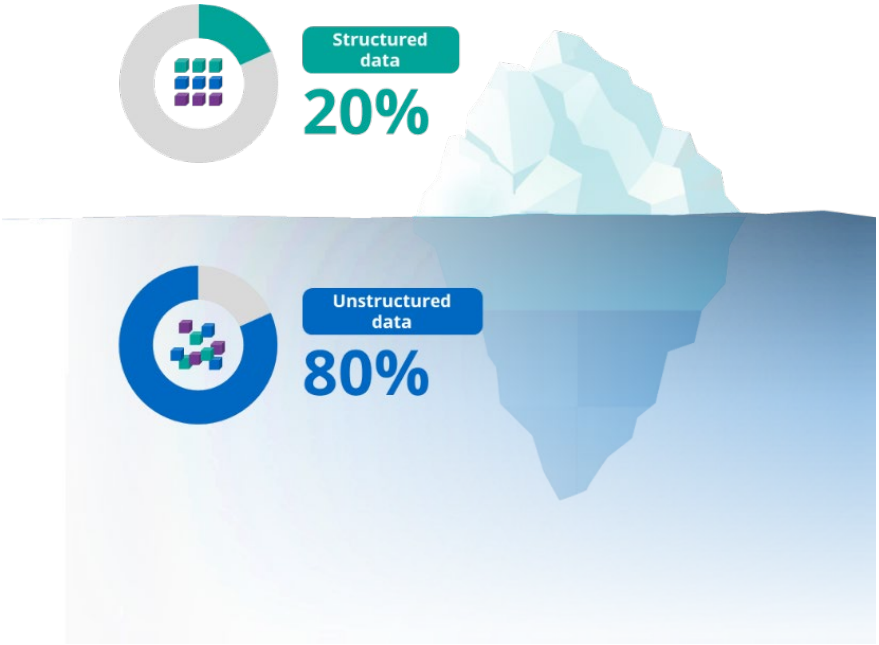
- Batch and real-time/ streaming
- Structured and unstructured

As well as three stages of data transformation:

OGL National Composite Centre 2022
This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

- Ingestion (curate, stage)
- Validate, clean, standardise and normalise
- Transform and enrich

Most data from intelligent manufacturing equipment would be classified as unstructured or semi-structured e.g., manufacturing data from IoT devices are often stored in a semi-structured format such as JavaScript Object Notation or JSON else could be images or sound files.



Traditional relational databases cannot be used to process this type of data efficiently and organisations typically require more storage for this type of data. Different analytic systems to those implementing traditional BI are required to process this type of data. Some of salient distinguishing characteristics between structured and unstructured data are given below.

Structured data

- Can be displayed in rows, columns and relational databases i.e. tabular
- Numbers, dates and strings
- Estimated 20% of enterprise data
- Requires less storage
- Easier to manage and protect with legacy systems

Unstructured data

- Cannot be displayed in rows, columns and relational databases i.e. non-tabular
- Sensor data, images, audio, video, emails, spreadsheets
- Estimated 80% of enterprise data
- Requires more storage
- More difficult to manage and protect with legacy systems

In general databases are used to store and manage information. Data in manufacturing have specific storage requirements, and therefore special needs regarding the type of database system utilised. The two types of databases required by manufacturing data are:

- Relational
- Non-relational or “NoSQL” (not only SQL)

OGL National Composite Centre 2022
This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

The following diagrams outline some details regarding these two technologies. Important to note that both relational and NoSQL DBs have their place – it is not a case of either/ or:

- Relational DBs – good at consistency (ACID)
- NoSQL DBs – good at performance/ reliability (uptime) – not good at consistency

Relational databases make sense in environments utilising structured data, whereas NoSQL databases makes sense when utilising unstructured data for Big Data processing, IoT, content management (online shopping) or data caching.

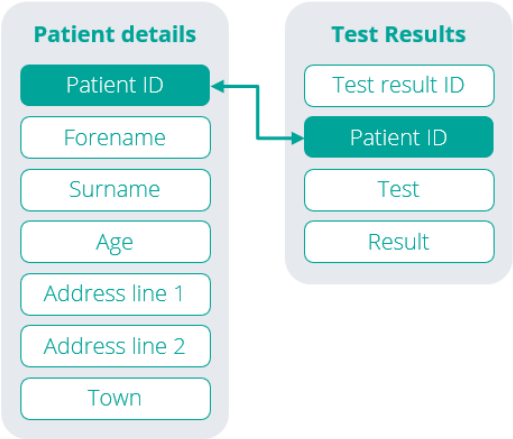
Relational databases were Invented by Edgar Codd (1970) and have the following properties:

- Database / schema defines entire collection of data
- Collection of tables make up a database
- Tables / entities contain information regarding specific "entity"
- Table contains records (rows) and fields (columns) which defines information about "entity"
- Relationships between tables (entities) using primary and foreign keys

Patient data and test results for a patient can be stored in a database in the following manner:

Patient ID	Forename	Surname	Age	Address line 1	Address line 2	Town
1234	John	Freeman	43	123 Main Road	Townsend	Heighton
5678	Anna	Kline	55	89a Orchard Lane	Old Oak	Heighton

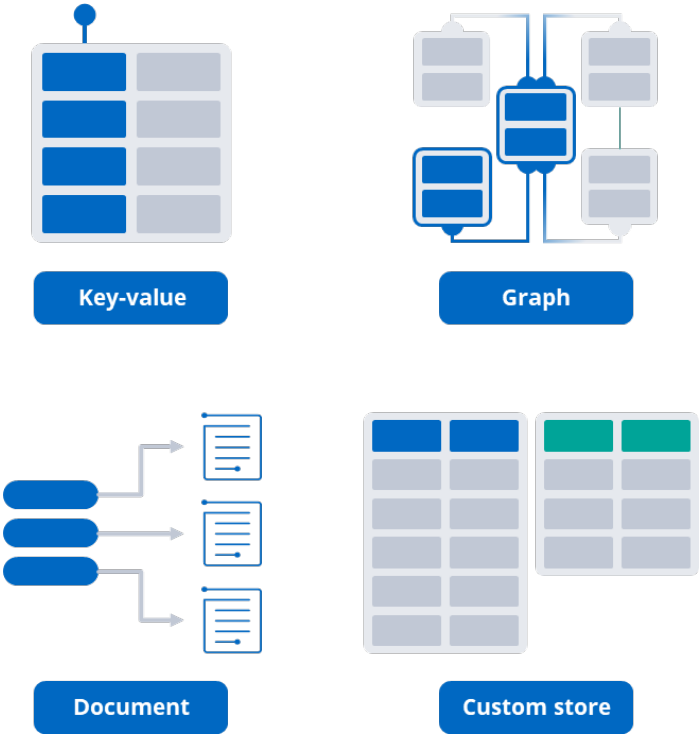
Test result ID	Patient ID	Test	Result
000053	1234	Pulse	75
000054	1234	BP	105/76
000055	1234	Respiration	45
000056	5678	Pulse	81
000057	5678	Respiration	50



[What is a Relational Database?]

NoSQL databases were designed to deal with large volumes of data in structured, semi-structured and unstructured format. Various types of NoSQL databases exist, each with its own pros and cons. The type of database used depends on your use case.

OGL National Composite Centre 2022
This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0



NoSQL Databases

For NoSQL databases, performance scales horizontally or "out", enabling it to adapt to big data application requirements. These databases are typically non-relational and schema-free and require programming skills to use. They typically don't enforce ACID properties, but BASE properties (eventually consistent).

To complement the abovementioned database technologies, cloud-based data management solutions typically include mechanisms to address:

- Management of DBMS systems on behalf of clients (configuration, patching etc.)
- Data ingestion (real time and batch, structured and unstructured)
- Data transformation (at scale)
- Data analytics (at scale)
- Information management and presentation
- Security

2) Data management in manufacturing

"Data Management is the development, execution, and supervision of plans, policies, programs, and practices that deliver, control, protect, and enhance the value of data and information assets throughout their lifecycles" (International, 2009).

Processes such as governance, accessibility, data literacy and a culture of experimentation should be encouraged. Simply having and storing data is not sufficient for data to generate value for an organisation. Typical goals of data management are as follows (International, 2009):

OGL National Composite Centre 2022
This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

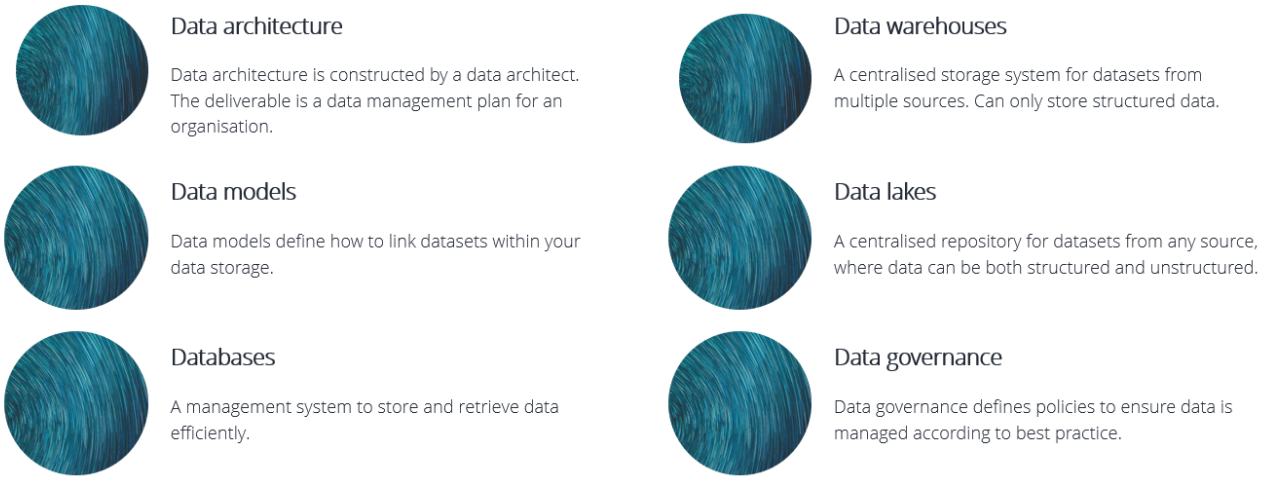
- Understanding and supporting the information needs of the enterprise and its stakeholders, including customers, employees, and business partners.
- Capturing, storing, protecting, and ensuring the integrity of data assets.
- Ensuring the quality of data and information.
- Ensuring the privacy and confidentiality of stakeholder data.
- Preventing unauthorized or inappropriate access, manipulation, or use of data and information.
- Ensuring data can be used effectively to add value to the enterprise.

The Global Data Management Community (DAMA) defines the various elements of data management as being:

- Data modelling and design
- Data storage and operations
- Data security
- Data integration and interoperability
- Document and content management
- Reference and master data
- Data warehousing and business intelligence
- Metadata
- Data quality
- Data architecture

DAMA.Org (requires registration)

For this course we will consider the following elements of data management selected from the above:



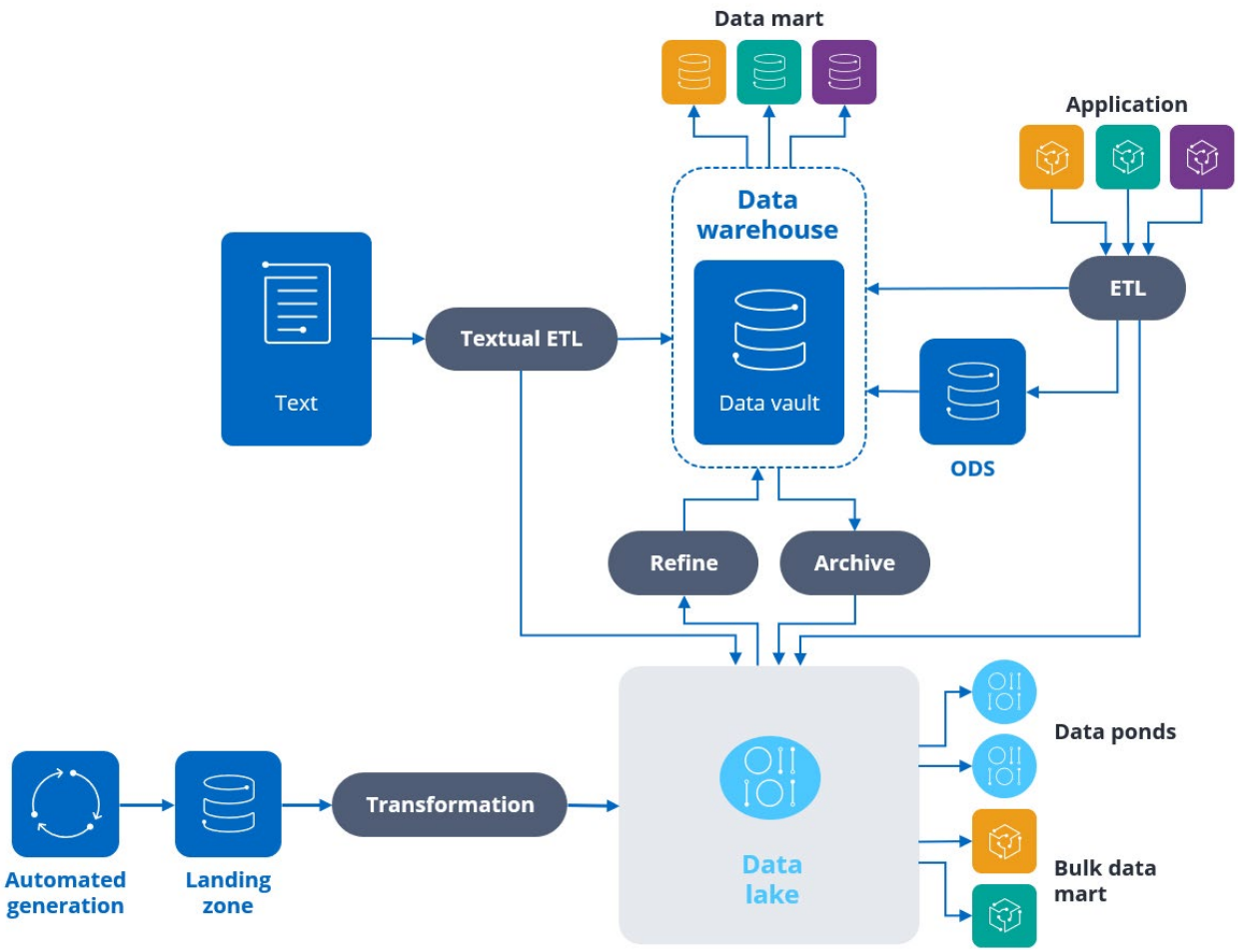
Data architecture for end-to-end manufacturing problems requires both structured and unstructured processing capabilities:

- Data Warehouses are tried and tested systems for hosting structured data
- Data Lakes can store and process unstructured data

OGL National Composite Centre 2022
This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

- It therefore makes sense to use both approaches for this use case

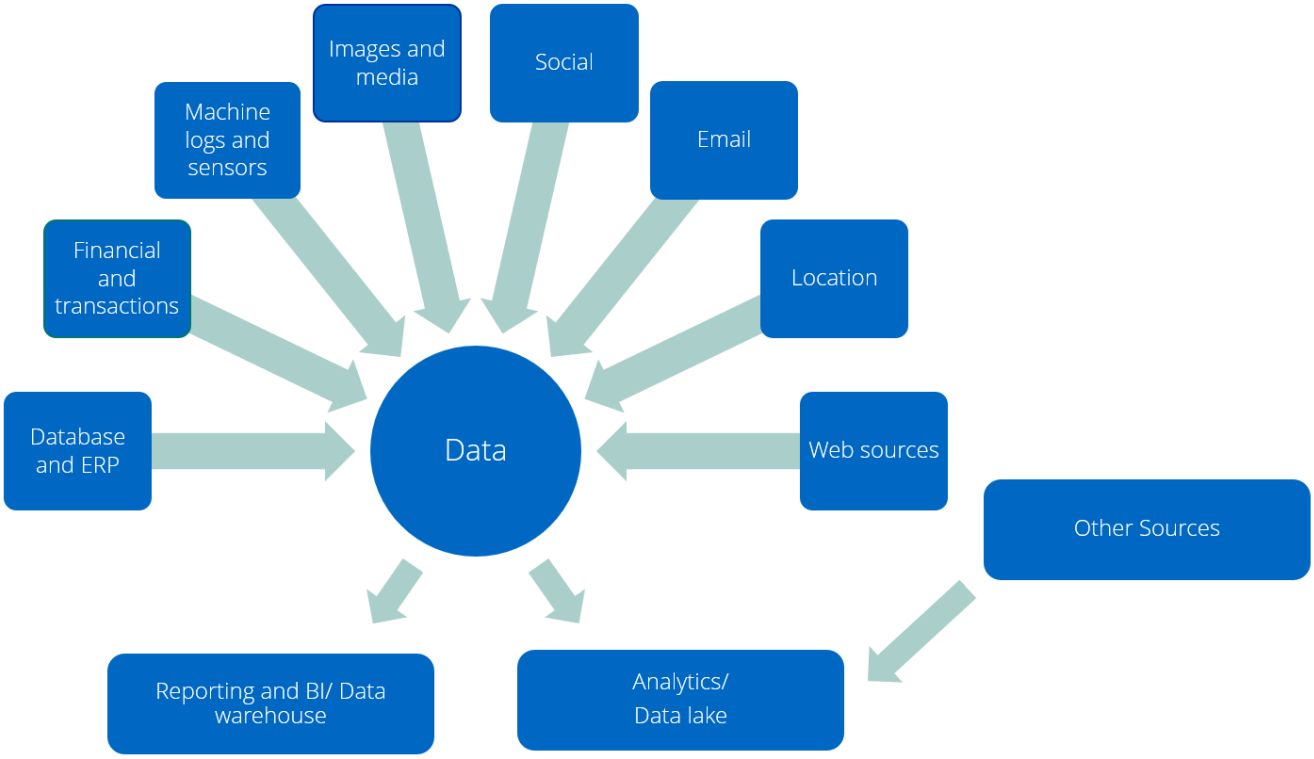
The following architecture makes use of both data warehouse and data lake architectures (Inmon & Linstedt, 2015):



There is no one right architecture, depending on resources available and use cases one can implement different approaches, provided best practices are followed. Top part of diagram shows a data warehouse, lower part a data lake. The data warehouse performs reporting of its own but also feeds into the data lake for additional value generation by integrating information with the lake. We will cover these concepts later in this section.

The figure below indicates a variety of different data sources and possible data management systems. See if you can identify the best suited data management system for each data source (**tip**: there is no one right answer, but there are some clear preferences for each source).

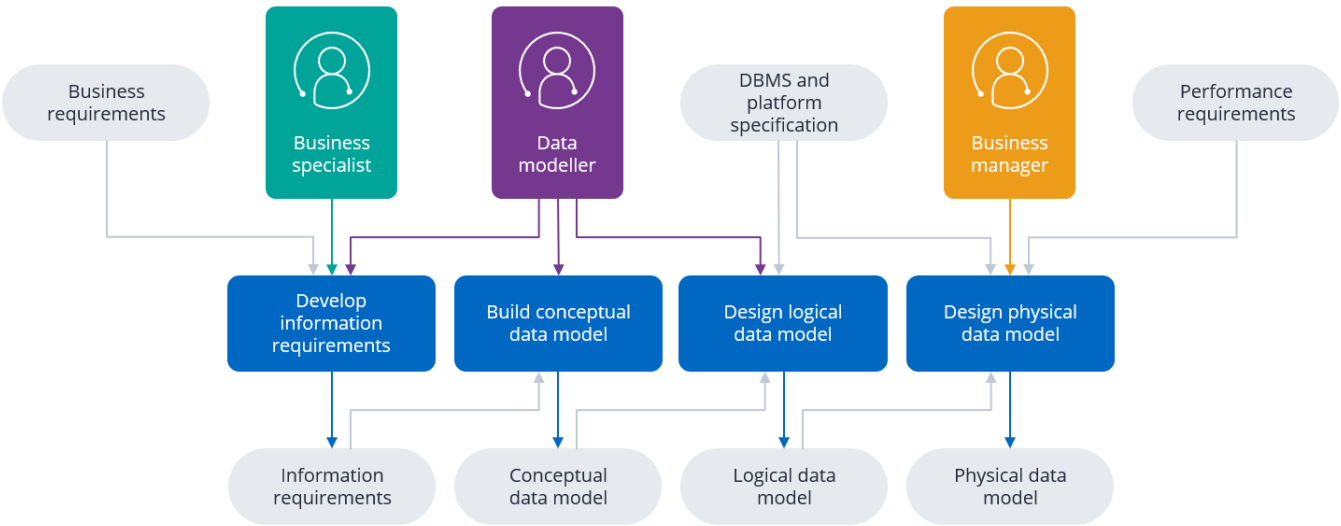
OGL National Composite Centre 2022
This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0



Data models help us to think about the problem that we are solving. This is arguably the most important part of building a data management system but often overlooked. We will next provide a few examples of outputs from a data modelling process. Three typical outputs from a data modelling exercise are a:

- Conceptual model
- Logical model
- Physical model

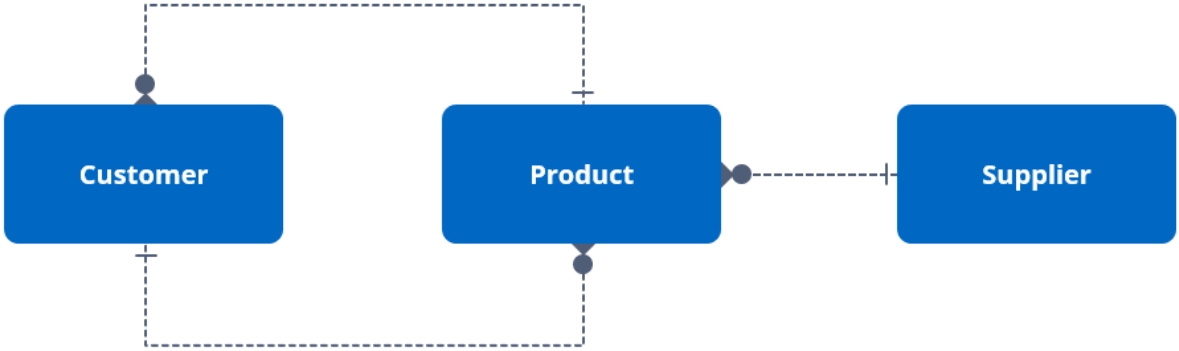
The modelling process in general is illustrated below (Detrano et al., 1989).



In practice the data modelling process is a lot less formal and is performed in its entirety by the data scientist/ engineer. This can cause sub-optimal data models (at all levels) but is a pragmatic approach that should be catered for.

A conceptual model in the manufacturing space could look as follows (Simsion & Witt, 2005):

OGL National Composite Centre 2022
This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

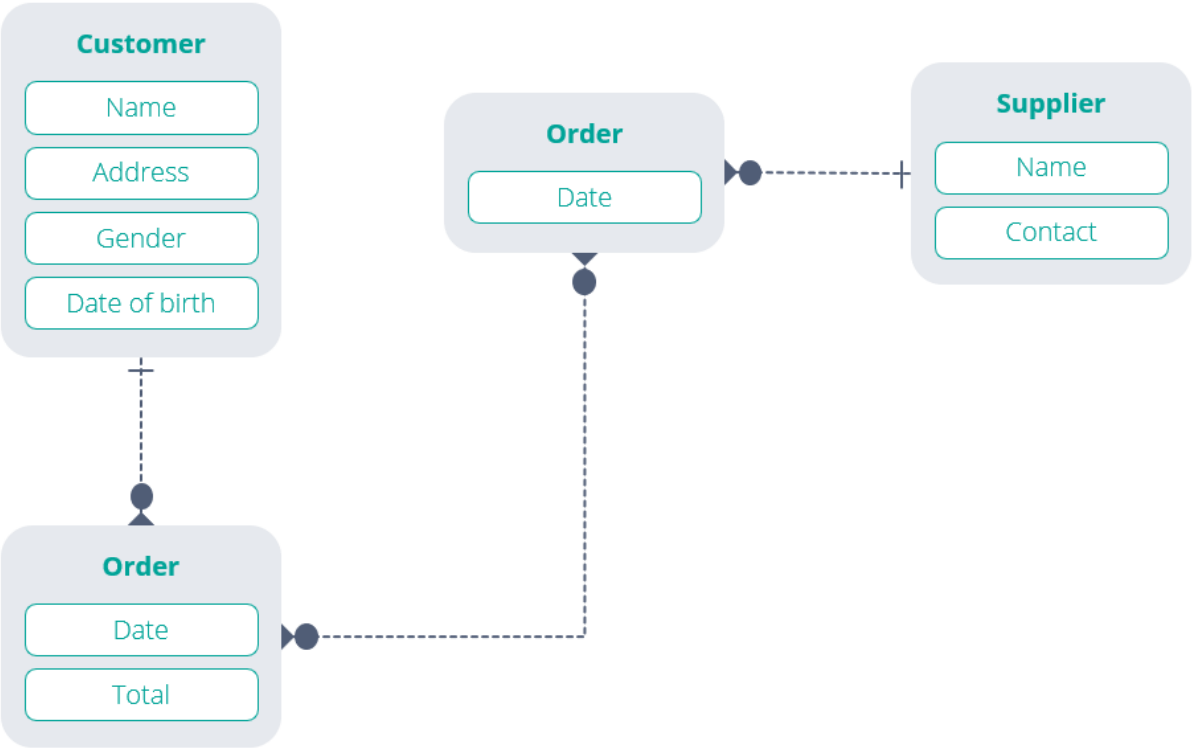


Conceptual modelling consists of a few activities:

- Elicitation of requirements
- Design of solution
- Evaluation of solution
- Definition of entities and relationships

A conceptual model should aim to define business entities and relationships between entities irrespective of the underlying database technology that will be utilised to implement the model. The conceptual model is typically generated using Unified Modelling Language (UML) Class Diagrams. This is the most overlooked component of data modelling. Most organisations skip the conceptual model and go straight to the logical model. This typically happens when software engineers or data scientists perform the modelling, rather than a business analyst (which should be the case).

The next level of modelling is the logical data model.

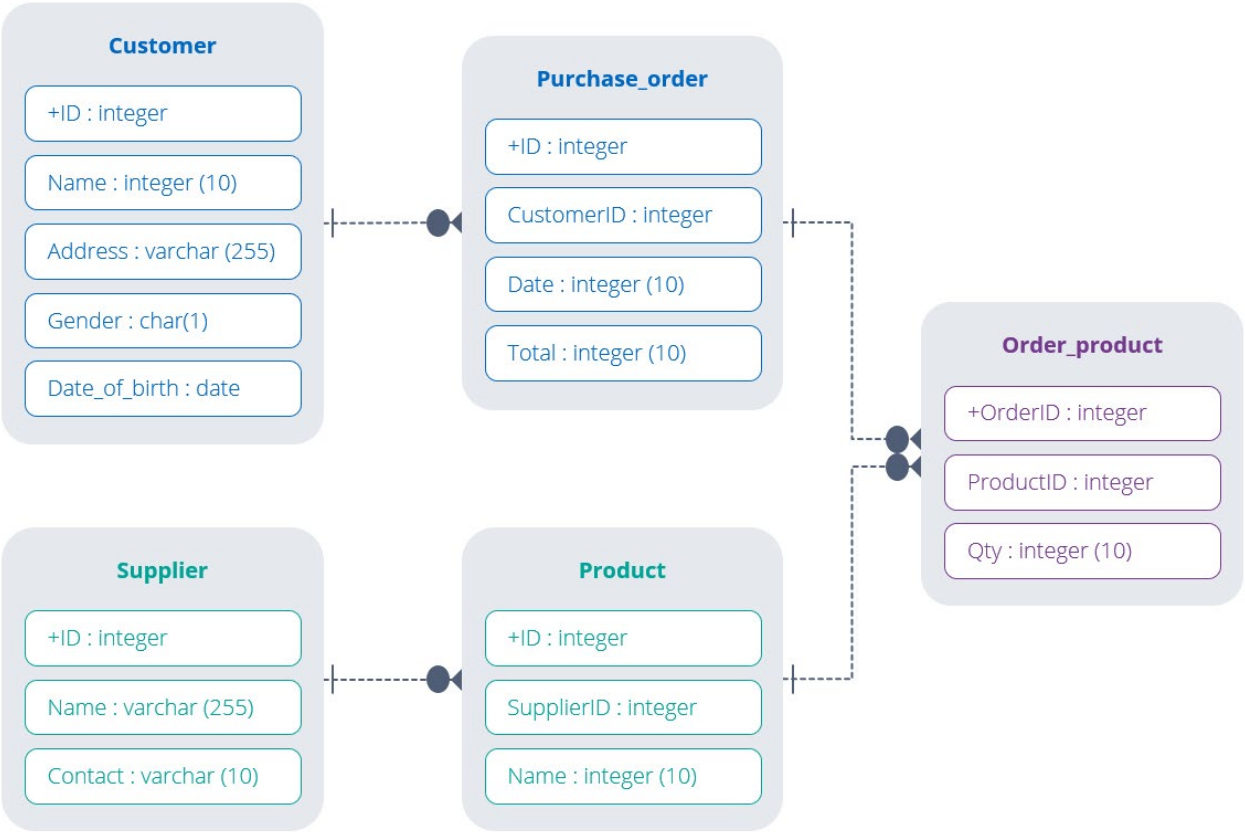


OGL National Composite Centre 2022
This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

At this level we decide on the type of database e.g., relational or NoSQL, but not the specific DBMS yet.

- Typically, logical models are for relational databases, but can be adapted for others i.e., NoSQL
- Computer Aided Software Engineering (CASE) tools can generate logical models from conceptual models
- Some typical steps for this stage are:
    - Implement relational relationships between tables based on conceptual model
    - Implement attributes
    - Specify primary and foreign key

The last level of modelling is the physical model.



At this point we make design decisions for a specific DBMS e.g., Apache Cassandra or PostgreSQL.

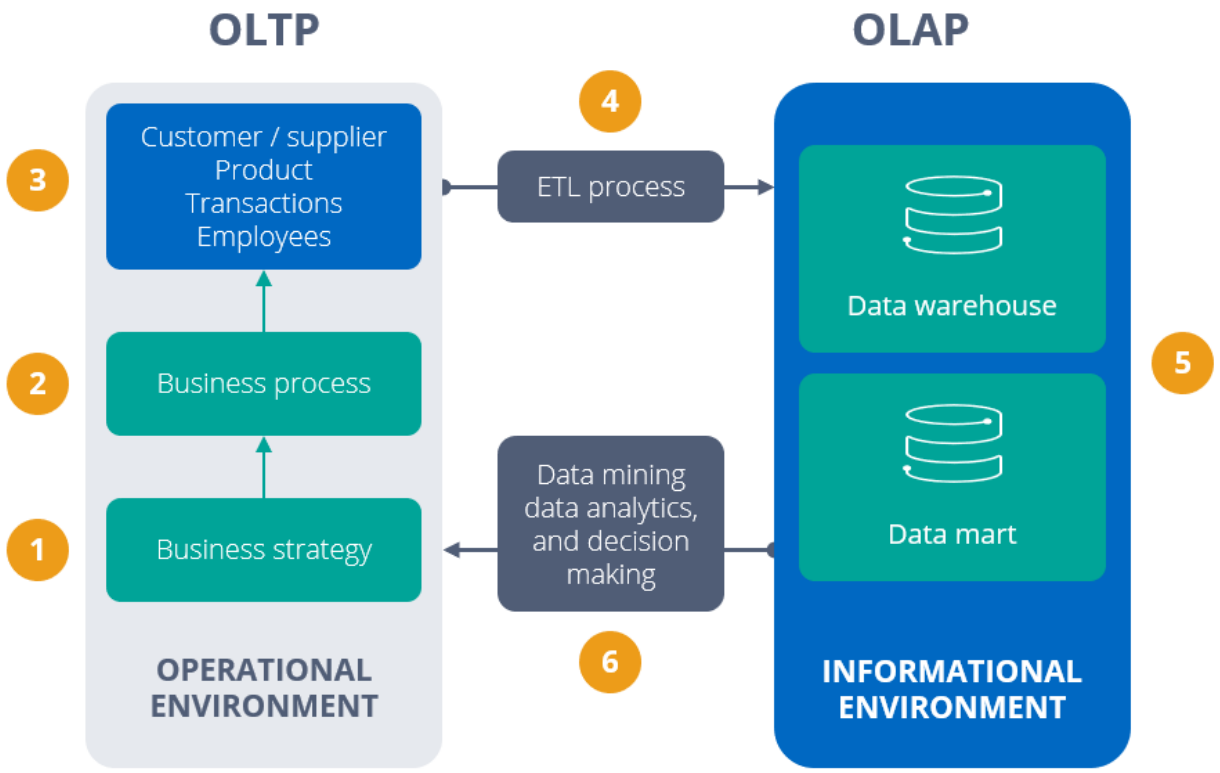
- The final database design is generated in terms of:
    - Tables, columns, and views
    - Indexes
    - Physical storage configuration
    - Data types

An important fact to know about is that for relational DBs, you need a different data model for operational/ transactional data vs informational/ analytic data. This is a significant fact that is often overlooked:

OGL National Composite Centre 2022
This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0
CATAPULT High Value Manufacturing

- For transactional systems – OLTP is used
- For analytic systems – OLAP is typically used

The following diagram illustrates this distinction.



Salient characteristics of both systems are as follows:

- OLTP's main operations are insert, update and delete (faster than OLAP)
- OLAP's main operation is to extract multidimensional data for analysis (faster than OLTP)
- The tables in OLTP database must be normalized (3NF)
- Tables in OLAP database may not be normalized

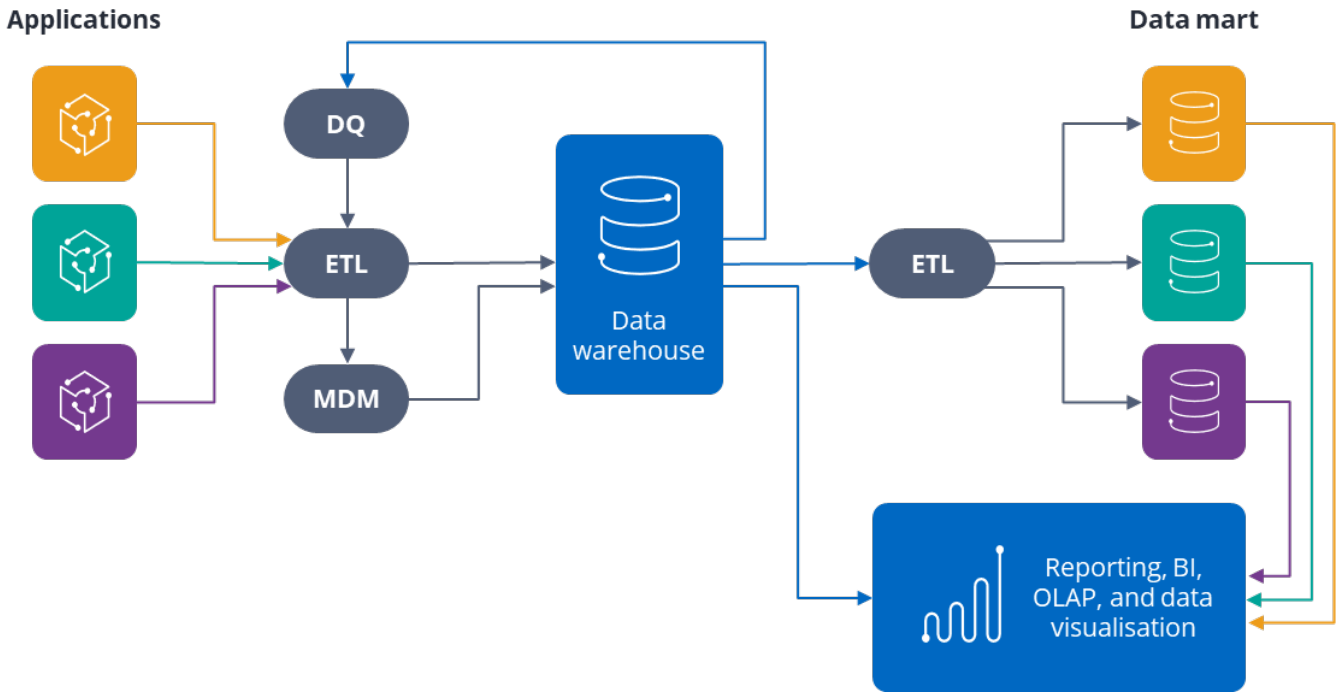
Taking the abovementioned into account you therefore need a DB for transactions and another one for analytics. Traditional BI organisations had an operational OLTP system, and an OLAP data warehouse for analytics. Since around 1990s, the intention to store all the data about an enterprise and make it available for analytics in a data warehouse was practised by most organisations. There were many failures, but overall, very this was a very successful architecture, and still is in many respects. Many technologies derive from data warehouse implementations including:

- Extract, transform, load (ETL)
- Data quality (DQ) and profiling tools
- Data modelling tools
- Business glossaries
- Metadata repositories

OGL National Composite Centre 2022
This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

- Data governance tools
- Master data management (MDM) systems
- Enterprise information integration (EII), data federation, and data virtualization tools

One significant drawback is that a data warehouse can only store structured data:



One important concept deriving from data warehouses is that of Master Data Management (MDM). This concept refers to the single trusted source that is stored in one location and everyone can trust the record to be true and accurate. While there are some practical considerations that means this is not always possible, it is a good goal to have and to strive towards. It can drastically reduce time spend searching, understanding, and processing data for use within your organisation.

Many data management technologies/ concepts derive from data warehouse implementations including:

- Extract, transform, load (ETL)
- Data quality (DQ) and profiling tools
- Data modelling tools
- Business glossaries
- Metadata repositories
- Data governance tools, data lineage
- Master data management (MDM) systems
- Enterprise information integration (EII), data federation, and data virtualization tools
- Structured data management systems

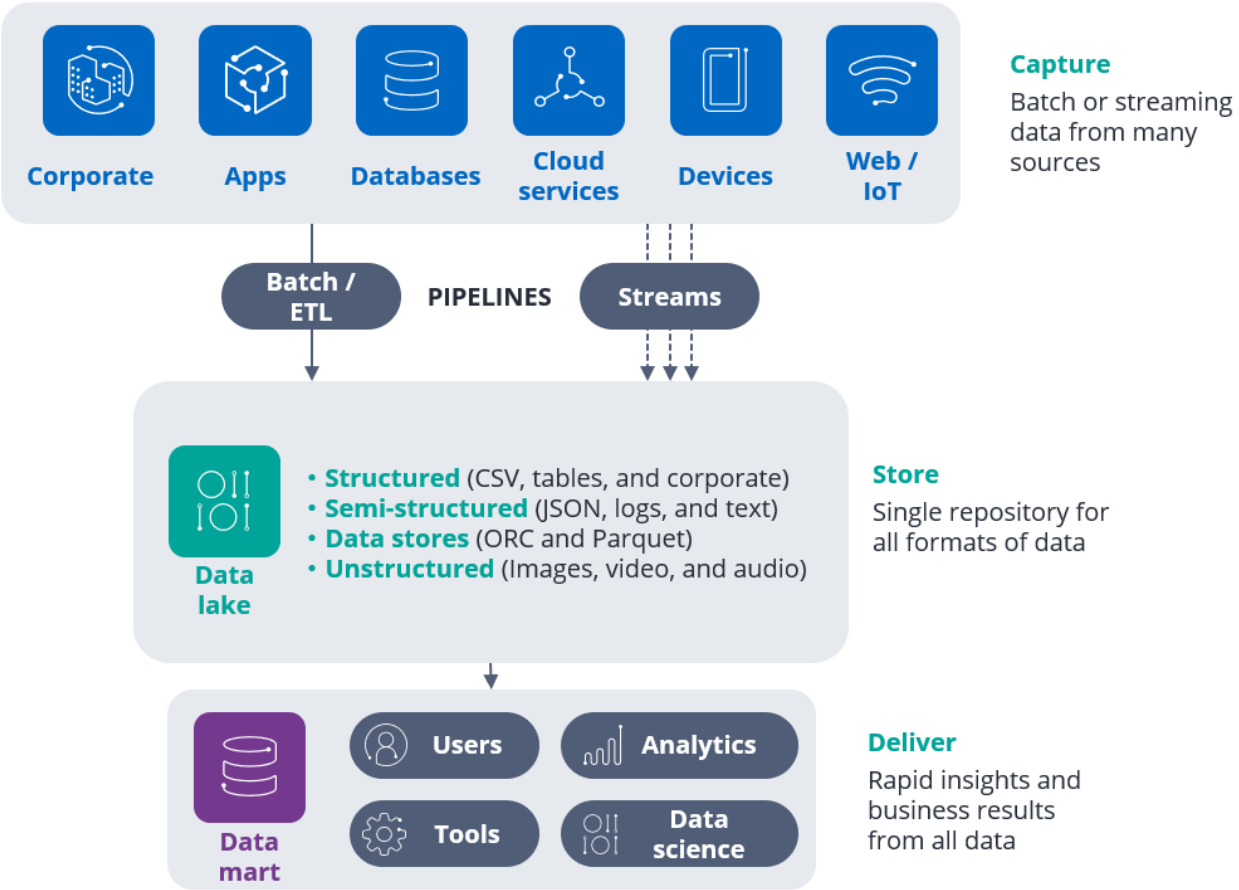
As opposed to a data warehouse, a data lake is a set of centralised repositories containing vast amounts of raw data (either structured or unstructured), described by metadata, organized into identifiable datasets, and available on demand. A data lake has the following characteristics:

- The data have no fixed schema and was consolidated from various sources
- All data formats should be possible (structured and unstructured)

OGL National Composite Centre 2022
This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

- The data have not been transformed i.e., is raw
- The data are conceptually present in one single place, but can be physically distributed
- The data are used by one or multiple experts in data science
- The data must be associated with a metadata catalog
- The data must be associated with rules and methods for their governance, data quality, data security and data lifecycle must be maintained

The figure below illustrates such a system.

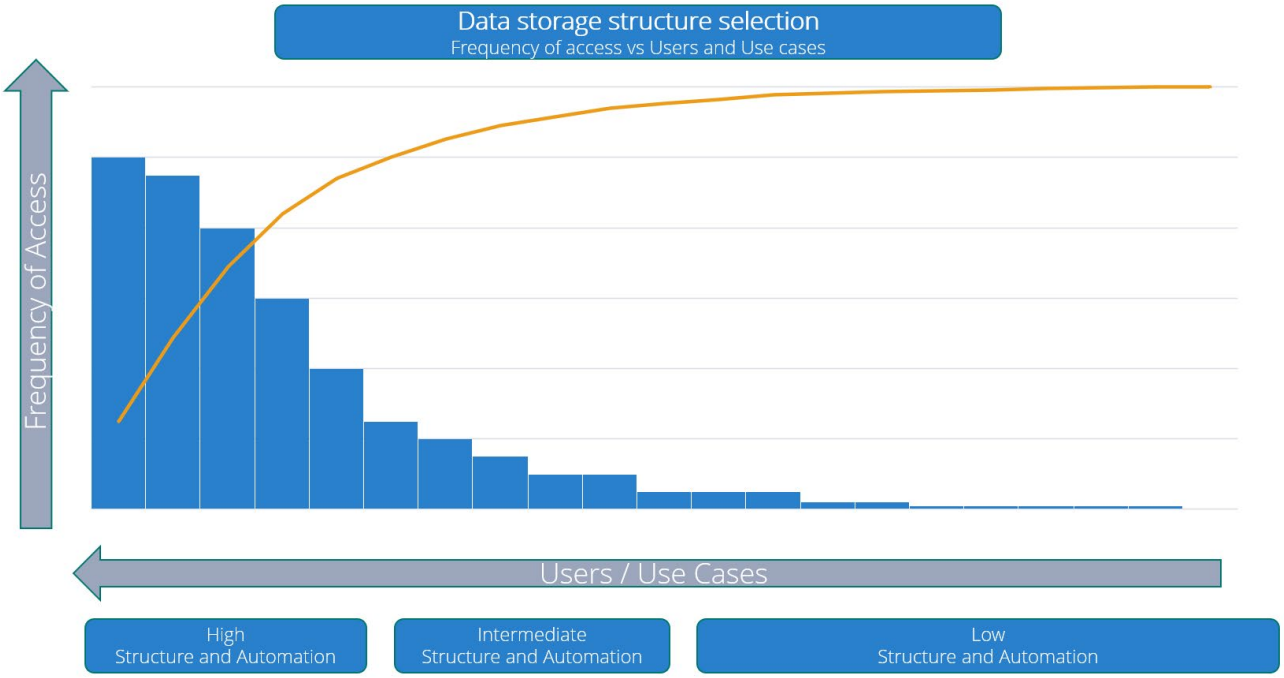


Metadata is key to the data lake architecture (in some ways it is the ***magic sauce*** of data lakes). The metadata catalog contains all the information a data lake needs to reconstitute information from the raw data stored in the lake. The data is only transformed when queried, hence why the process is called Extract/ Load/ Transform or ELT as opposed to Extract/ Transform/ Load or ETL which data warehouses use.

It is important to realise that different data types and, in some cases, different uses of similar data types require different storage structures. To some extent the management of data has become simpler with the adoption of cloud computing and the tools available within these environments for data management, but it is still important to understand how to select fit for purpose elements for your use case.

The following diagram poses the question of whether a spreadsheet, database, data warehouse or data lake would be the storage structure of choice for different use cases (spreadsheets to the right and data lakes/ data warehouses to the left):

OGL National Composite Centre 2022
This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0



The decisions that you make regarding storage and processing have a significant impact on your ability to utilise the data in your organisation. We briefly alluded to unstructured data as data that has not been processed to a useful format. You will find that there is a core set of data (such as financial or operational data) that is used to manage the core activities of your organisation. When you start looking at additional data sources these sources may be of interest to specific departments or users, and we often find that these sources do not receive the same amount of effort in processing and storing the data. In cases where there is a once-off or exploratory project and the data is seen to be of value, it may be worthwhile incorporating this in your main sources of data as more and more users become interested in this set of data. We often see a migration of data sources from unstructured and unmanaged based on the frequency of use and number of users interested in the data rather than overly ambitious projects aiming to standardise all the data in your organisation. Data warehouses are an example for formal structures. While there is significant effort and costs involved in planning, creating, and maintaining data warehouses, they offer the ability for a large portion of users to access standardised data using tools such as SQL, or graphical tools to visualise data without having to source and prepare the data themselves. Data that has not been standardised and processed are typically retained in more informal areas such as data lakes, databases, or spreadsheets, but due to the data not being as well understood and processed, a higher level of skill and different tools are required to access this type of data storage. Viewed from another perspective, if your report or analysis is ad-hoc or very infrequent, it makes sense not to add this to your roadmap of elements to be added to more formal structures.

Note that the type of data as well as the usage thereof would have a significant impact on the storage structure that is selected. Once you move away from spreadsheets and local files you would typically start to look at databases. From an analyst, data miner, or data scientist perspective you need to be aware of the different types of storage structures, but do not necessarily be an expert in all the technologies. Make sure to understand the different types of states (streaming or at rest), performance and security expectations, pre-processing required and ease of use.

Sources of data can include internal data generated by machines or processes, manually captured items or existing information systems. Data can be subjective or objective and we have concepts such as metadata which is data about data. We also often supplement internal data with external

OGL National Composite Centre 2022
This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

sources. These can be used to expand the usefulness of internal processes or can be used to test the viability of a new approach prior to committing the resources to generate the data internally.

Not all data has the same value, where some data is frequently used for multiple use cases and by multiple users. These data sets often require high levels of accuracy and are processed using standardised and scheduled processes. Think of financial reports as an example of high value, frequently used data. In other cases, you may be unsure about the quality and value of data. You may have a new or not currently used source of data where you first need to establish the value of the dataset before deciding whether to make this available for use by other stakeholders in your organisation.

From an analytical point of view, you need to make informed decisions, and inform your downstream users of the risks associated with using data of various qualities in your analysis. As we go through the process of exploring the data, trying to identify or model the underlying signal and using these insights to predict or classify or otherwise informing decisions within the organisation, we need to work with the specific questions and datasets to determine what is acceptable or not. If we were to have an arbitrary number of corrupted or missing records in the case of the finance department it would be a significant problem. However, if you are performing exploratory analysis on sensor or other process data where you have access to significant amounts of data it might not be a problem at all, as a single value by itself doesn't carry much information. In the latter case we are interested in the patterns and behaviour rather than the individual observations. If the same data is used for a highly sensitive production model, it may be a problem, but you would have to perform additional work to determine whether this is the case or not. Make sure to think critically about your needs, that you understand the data, that you document your assumptions (and business rules if you apply transformations), and very importantly, that you manage the expectations of your downstream stakeholders by acknowledging the risks associated with your data sources. Other users or departments may have different expectations around stability and availability of data that you will need to account for. Are you trying to create a model that describes general behaviour, or counting specific units?

Suggested reading

- KDNuggets - Asking 10 questions
- BMC: Data lakes vs Data Warehouses

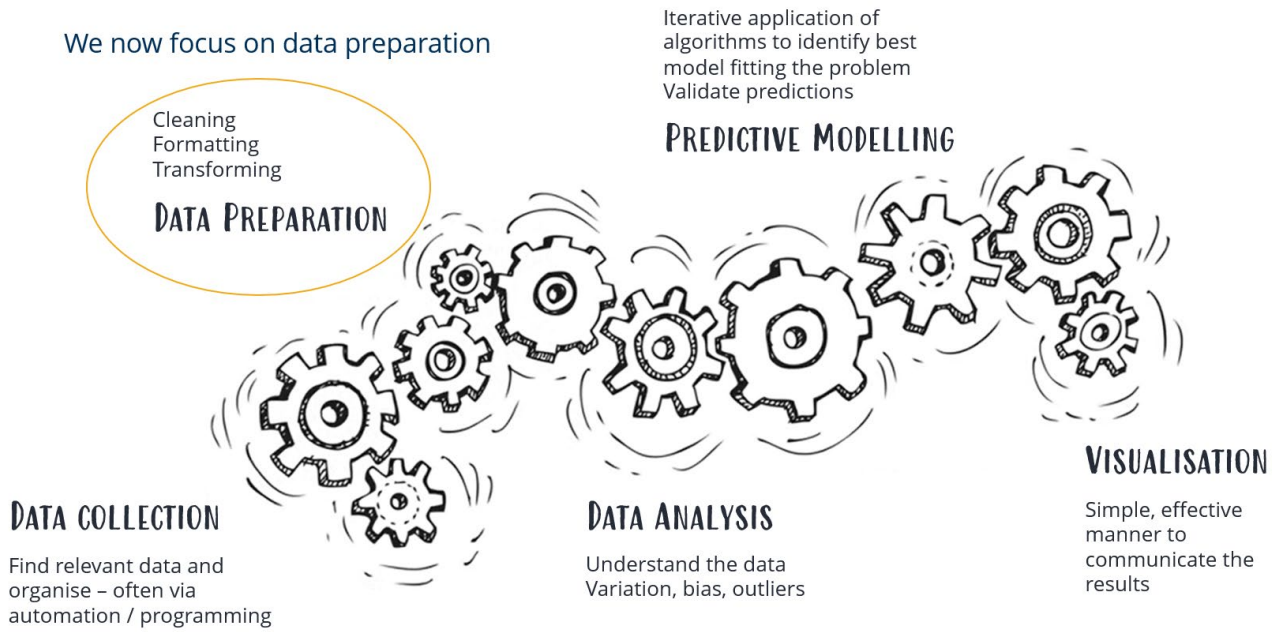
Based on the articles

- Think about data sources and use cases in your workplace.
- How critical is the integrity of the data?
- Are there risks that need to be communicated with stakeholders?

3) Data preparation

We now move on to the next topic in data management which is "data preparation" or "explore/ clean/ transform". Recall the workflow presented in unit 1?

OGL National Composite Centre 2022
This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0



Data need to be prepared for ML for the following reasons:

- Machine learning algorithms require data to be numbers
- Some machine learning algorithms impose requirements on the data
- Statistical noise and errors in the data may need to be corrected
- Complex nonlinear relationships may be teased out of the data

To apply the statistical concepts to unstructured raw data we must therefore process and manipulate the data into a structured form. This structured form typically takes on one of the following forms:

Continuous: These are data elements that are defined on an interval that can be limited or unlimited. Examples include income or sales.

Categorical:

- *Nominal:* These are data elements that can only take on a limited set of values with no meaningful ordering in between. Examples include marital status, profession, purpose of loan.
- *Ordinal:* These are data elements that can only take on a limited set of values with a meaningful ordering in between. Examples include credit rating; age coded as young, middle aged, and old.
- *Binary:* These are data elements that can only take on two values. Examples include gender, true or false.

This step in a data analytics project is generally referred to as data preparation, although it goes by many other names, such as *data wrangling*, *data cleaning*, *data pre-processing* and *feature engineering*.

Before starting with data preparation, it is useful to consider the following questions first:

Data governance questions:

- Who is the data owner?
- Are there data governance policies in place for the dataset?
- Any relevant regulation and legislation that should be considered?

Data documentation:

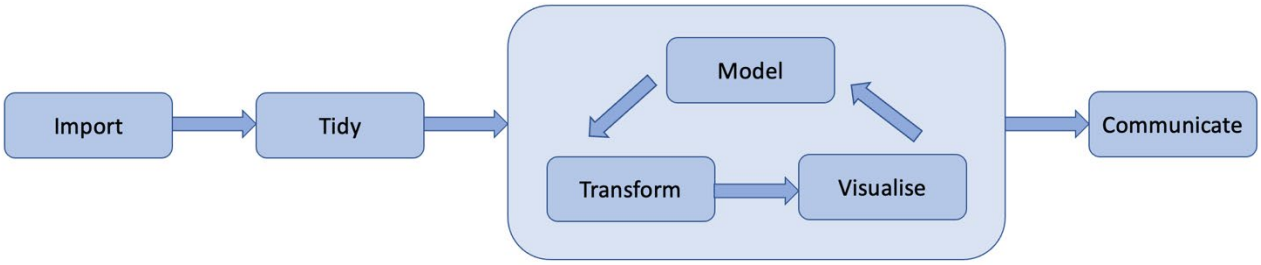
- Is there metadata or a data dictionary available?
- Is there a change management process in place for the data?
- What was the context of collection for the data?

Understanding the data:

OGL National Composite Centre 2022
This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

- Complete and accurate (independence, distribution, bias)
- Data distribution – need for standardisation?
- Dealing with errors and outliers
- Data quality

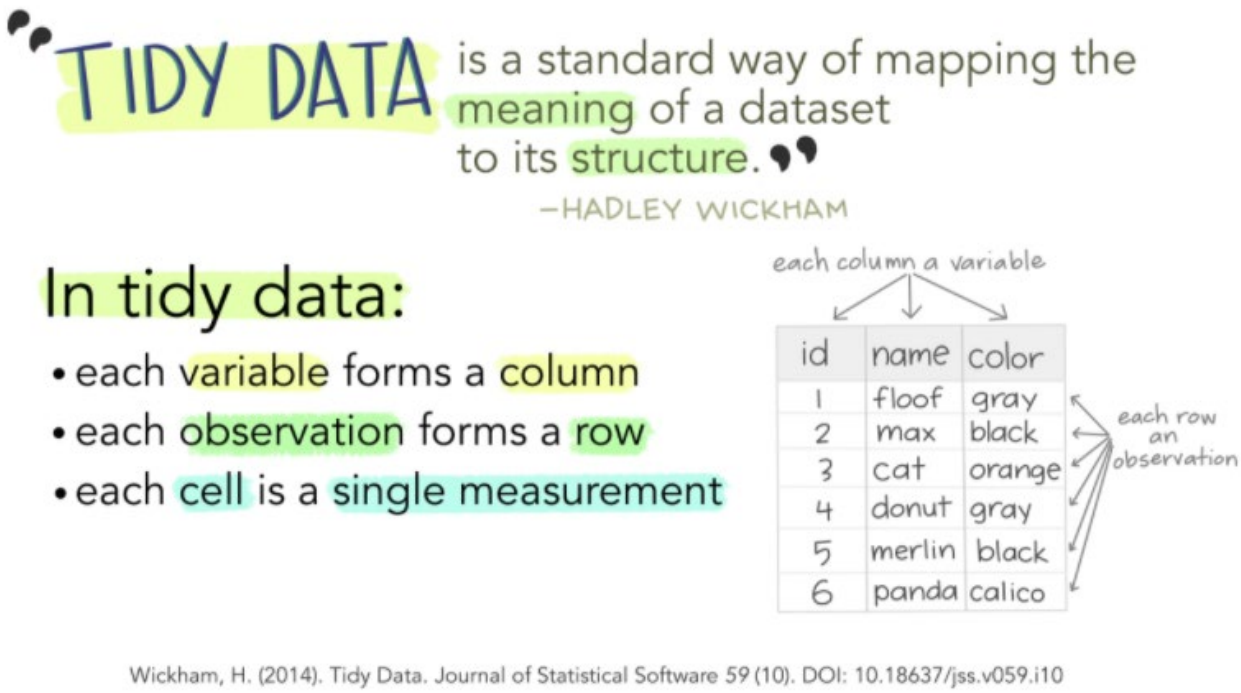
Data preparation often takes two forms, general steps to clean up data in for further use, and specific transformations required for the use case you are working on. We will start by considering cleaning or tidying of the data.



Suggested reading

- Tidy Data

In the suggested reading, make sure to note the three fundamentals of tidy data as described by Hadley Wickham in his 2014 paper.



Tidy Data

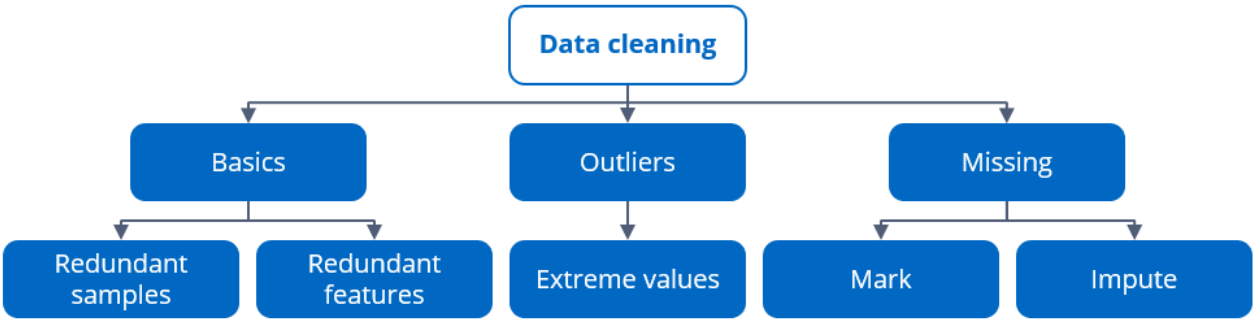
Do not underestimate the amount of effort and resources required to perform these steps. You should also note that while it is often stated that these steps often take up to 80 or 90 percent of the effort spent, there are also significant opportunities to better understand the processes that generate, process and store data for your particular use case. In many cases the insights gained during the data exploration process are of significant value to the organisation and it is important to have processes in place to capture and action these insights. As example, insights gained in ad-hoc analysis may contain useful rules for data warehouse or business intelligence teams.

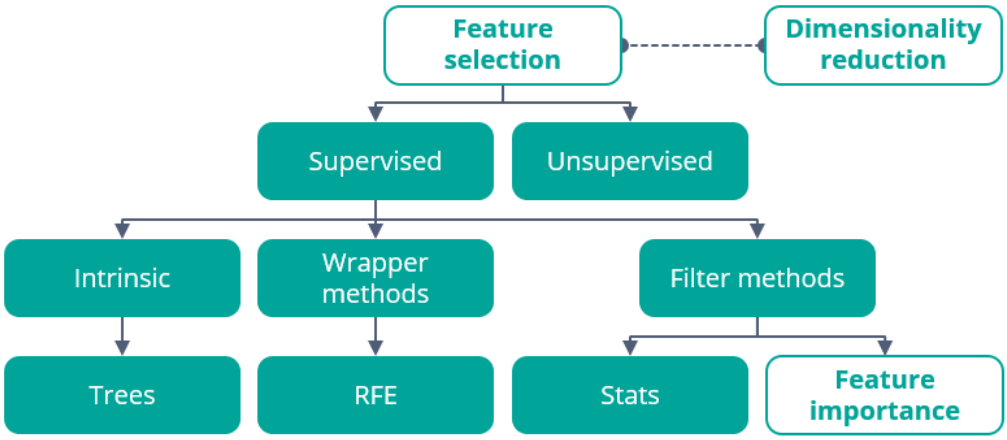
OGL National Composite Centre 2022
This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

The next steps depend on your analysis, but typically involve the following elements:

- **Data Cleaning**: Identifying and correcting mistakes or errors in the data
- **Feature Selection:** Identifying those input variables that are most relevant to the task
- **Data Transforms:** Changing the scale or distribution of variables
- **Feature Engineering**: Deriving new variables from available data
- **Dimensionality Reduction:** Creating compact projections of the data

The following are steps that can generally be followed to clean select features:





Discussion

Reflect on the material presented and try to identify a few different data sources from your own business or an imaginary use case:

- What type of data is involved?
- Who would be interested in this data?
- What could it be used for? (opportunities)
- What are the dangers involved in using this dataset? (risks)
- Should the data be stored in more formal structures or not?
- What are the suggestions regarding tidying the data and preparing it for your selected use case?
- Who should have access to the data?
- You can also think about the data in terms of different types of questions that would be of interest to your organisation at different phases of maturity.
- What happened?
- What is happening?
- What is going to happen?
- How can I make it happen?

OGL National Composite Centre 2022
This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

4) Supplemental material

- *KDNuggets: List of data sources*
- *Data Architecture: A Primer for the Data Scientist -- Big Data, Data Warehouse and Data Vault*

OGL National Composite Centre 2022
This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

5) References

Bajic, B., Rikalovic, A., Suzic, N., & Piuri, V. (2021). Industry 4.0 Implementation Challenges and Opportunities: A Managerial Perspective. *IEEE Systems Journal*, *15*(1), 546–559. https://doi.org/10.1109/JSYST.2020.3023041

Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J. J., Sandhu, S., Guppy, K. H., Lee, S., & Froelicher, V. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. *The American Journal of Cardiology*, *64*(5), 304–310. https://doi.org/10.1016/0002-9149(89)90524-9

Inmon, W. H., & Linstedt, D. (2015). *Data Architecture: A Primer for the Data Scientist -- Big Data, Data Warehouse and Data Vault*. Morgan Kaufmann. http://www.sciencedirect.com/science/book/9780128020449

International, D. (2009). *The DAMA Guide to the Data Management Body of Knowledge - DAMA-DMBOK*. Technics Publications, LLC.

Simsion, G. C., & Witt, G. C. (2005). *Data modeling essentials* (3rd ed.). Amsterdam; Boston.

Kleppmann, M. (2017). Designing Data-Intensive Applications: The Big Ideas behind Reliable, Scalable, and Maintainable Systems. In *O'Reilly Media, Inc.* O'Reilly.

Gorelik, A. (2019). *The Enterprise Big Data Lake*. O'Reilly.

Zheng, A., & Casari, A. (2018). *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists*. O'Reilly.

Torres, M. T., Hart, G., & Emery, T. (2019). *The Dstl Biscuit Book Artificial Intelligence, Data Science and (mostly) Machine Learning*. 1–40.

OGL National Composite Centre 2022
This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

6) List of further courses

Course	Topic	Level
Become a Data Engineer – Udacity	Learn to design data models, build data warehouses and data lakes, automate data pipelines, and work with massive datasets. At the end of the program, you'll combine your new skills by completing a capstone project.	Intermediate
IBM Data Engineering Professional Certificate - Coursera	You will learn to use the Python programming language and Linux/UNIX shell scripts to extract, transform and load (ETL) data. You will work with Relational Databases (RDBMS) and query data using SQL statements. You will use NoSQL databases and unstructured data. You will be introduced to Big Data and work with Big Data engines like Hadoop and Spark. You will gain experience with creating Data Warehouses and utilize Business Intelligence tools to analyse and extract insights.	Beginner

OGL National Composite Centre 2022
This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0

7) Copyright notices

Various copyrights utilised in these lecture notes:

Reference	Copyright
The DSTL Biscuit Book	http://www.nationalarchives.gov.uk/doc/open-government-licence/version/3
Tidy Data	https://creativecommons.org/licenses/by/4.0/

OGL National Composite Centre 2022
This IDT403 Introduction to Data Analysis and Machine Learning content is licensed under the Open Government Licence 3.0