Data Understanding and Data Modelling

An Introduction – Discover Data Science for Professionals track

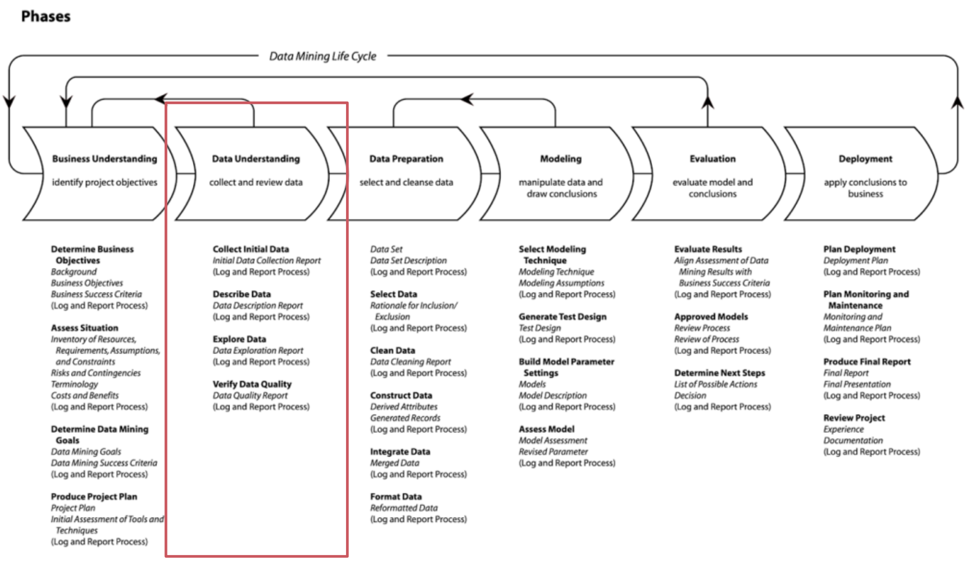
version April 2021

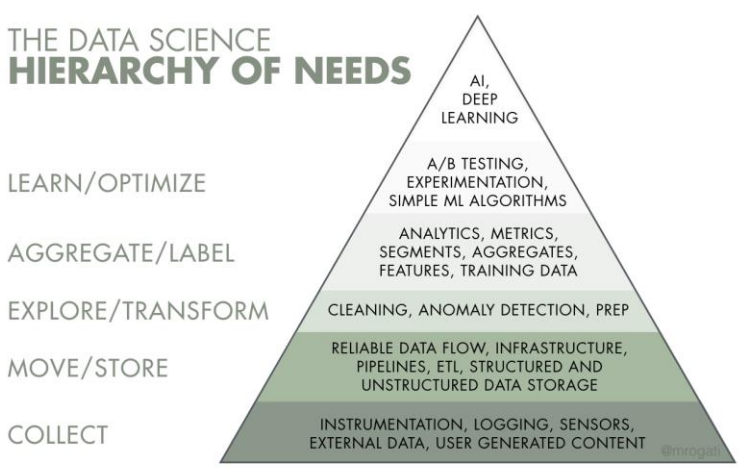
TU/e Eindhoven University of Technology · Tilburg University · 's-Hertogenbosch · Provincie Noord-Brabant · JADS Jheronimus Academy of Data Science

Copyright notice

This material is the intellectual property of Joran Lokkerbol, Daniel Kapitan and JADS. This material is provided to you for personal use only. Sharing, posting or selling is strictly forbidden without permission from the author.

Today’s agenda





Whatever the goal,

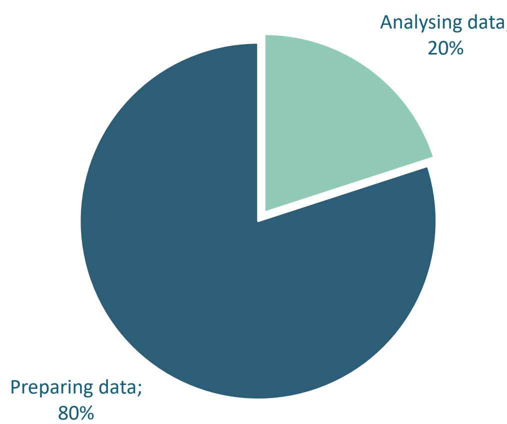
whatever type of analysis:

always start with a thorough description

and understanding of the data

https://hackernoon.com/the-ai-hierarchy-of-needs-18f111fcc007

Just to manage expectations:
what do data scientists actually do?



The smallest data science team

> If possible, try not to work on your own but at least with one buddy

> Make a distinction between data engineering and data science activities
  - Helps you to structure your workflow and pipeline
  - Separation of concerns

> More specialists roles are possible as your team grows
  - Data Consultant
  - Data visualization (front-end apps)
  - DevOps
  - …

https://www.datacamp.com/community/blog/data-scientist-vs-data-engineer

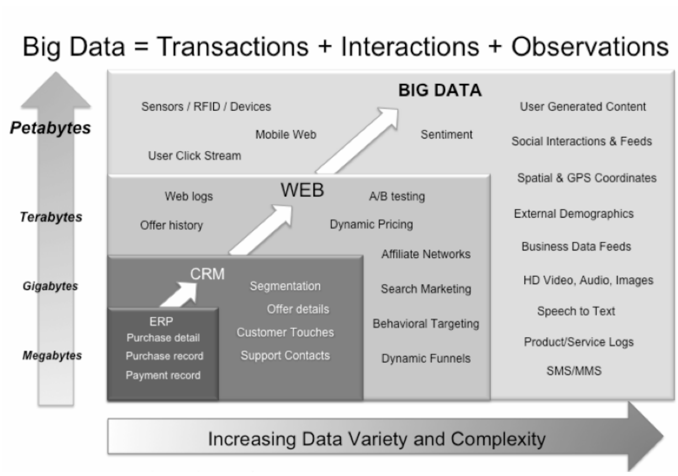
Agenda

section	topic
Data Engineering (very short introduction)	Data types
	Modern data architectures
	Data modeling
Data Understanding	Collect data
	Describe data
	Explore data
	Verify data quality

Data Engineering

A very short introduction

Evolution of enterprise data



https://www.researchgate.net/figure/Big-Data-Transactions-with-Interactions-and-Observations-Source_fig3_243963821

How does a computer store data?

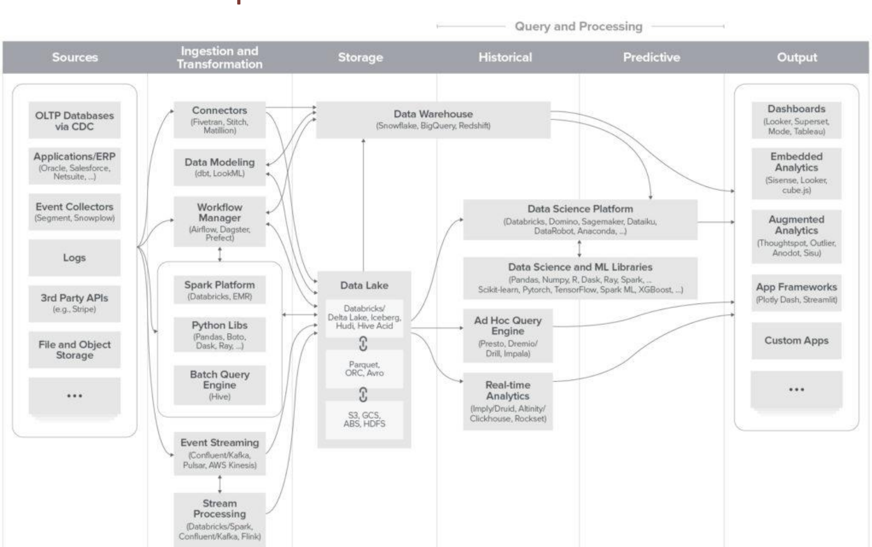
- Structured data: databases, data model
- Semi-structured data: json, XML, relationships with graph database
  - 'John is a friend of Mary'
- Unstructured data: audio files, images, video

Steven’s topology of measurement (1946)

scale	measure property	math operations	advanced operations	central tendency
nominal	classification, membership	=, ≠	grouping	mode
ordinal	comparison, level	>, <	sorting	median
interval	difference, affinity	+, -	yardstick	mean, deviation
ratio	magnitude, amount	x, /	ratio	geometric mean, variation

https://en.wikipedia.org/wiki/Level_of_measurement

The main components of a modern data infrastructure



https://a16z.com/2020/10/15/the-emerging-architectures-for-modern-data-infrastructure/

From conventional ETL to modern data pipelines

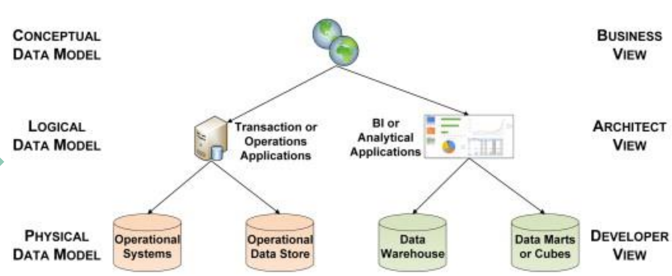
conventional ETL (Extract-Transform-Load) [1]	modern data pipelines [2]
processing of (semi-)structured data	processing of all kinds of data (incl. unstructured)
works with updates and transformations to save storage	follows principal of immutable data with more copies
more simple, linear process flows executed on central data warehouse	more complex, distributed process flows (directed acyclic graph) run on various machines
focuses on creating data marts and dashboard	many different use-cases, including machine learning, streaming processing

https://en.wikipedia.org/wiki/Extract,_transform,_load
https://en.wikipedia.org/wiki/Directed_acyclic_graph#Data_processing_networks

Three levels of data models

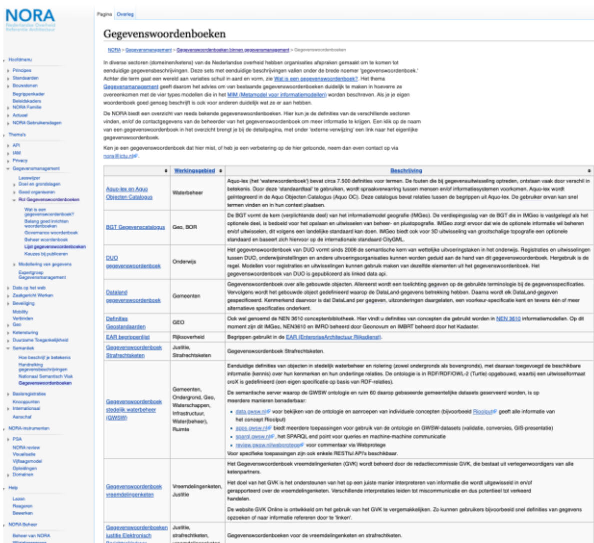
A data model is (Wikipedia):

An abstract model that organizes elements of data and standardizes how they relate to one another and to the properties of real-world entities.



https://www.sciencedirect.com/topics/computer-science/logical-data-model

Conceptual models and semantic standards



- Standardized definition of information elements *(informatiebouwstenen)*
- Over [30 standards](https://www.noraonline.nl/wiki/NORA_online) currently in use in the Netherlands

https://www.noraonline.nl/wiki/NORA_online

Example: conceptual data model Dutch national registries

https://www.stelselcatalogus.nl/stelselplaat

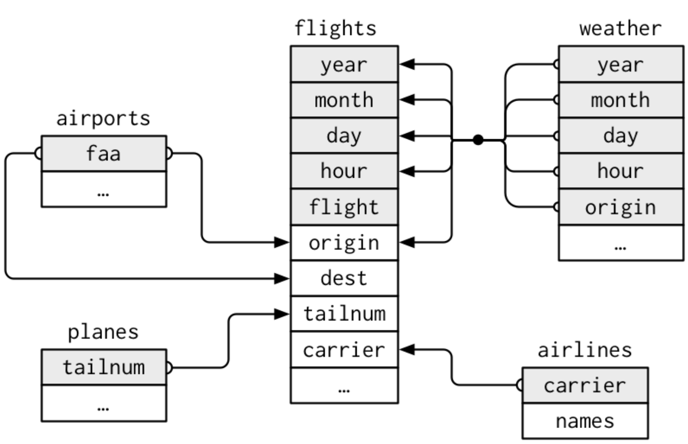
Three different logical data model types

> third normal form
  - used for processing transactions
  - optimized for fast inserts and updates

> data vault schema:
  - used in data vaults
  - optimized for flexibility for integrating different sources

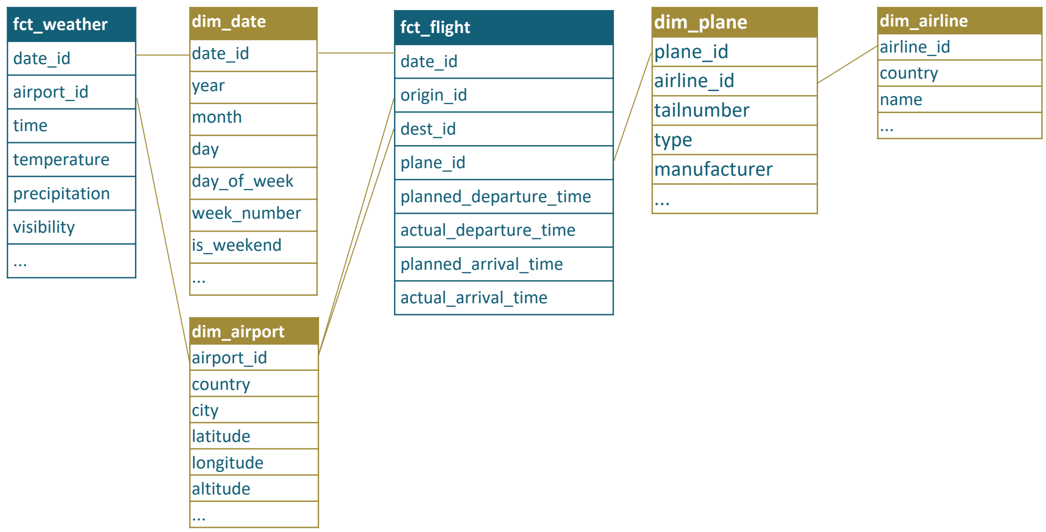
> star schema:
  - used in data marts
  - optimize of fast querying and online analytical processing (OLAP)

Example: original data in third normal form



https://r4ds.had.co.nz/relational-data.html

Example: data in star schema



Data Understanding

Collecting, describing, exploring and verifying your data

Evolution of enterprise data

Why bother with data understanding?

> It checks the feasibility of a project

> It contributes to refinement of the project scope

> It provides explicit input for next steps in your project, namely:

- How to clean data
- How to adjust existing variables
- Where to create new variables
- How to approach data modelling

Insight into the available data

> Create an overview of the available data
  - available variables
  - number of observations per variable
  - levels within categorical variables
  - descriptives of numeric variables

> Describe the population
  - How did observations (clients, events, etc) make it into the dataset?
  - What do descriptives say about the population?

> Create an overview of data domains that are relevant for your problem

Insight into data quality and usability

> Explore quality:
  - number of missings per input variable
  - number of missings in the outcome variable
  - occurrence of impossible values or combination of values
  - occurrence of outliers

> Explore usability
  - Are data available at the intended moment of prediction or classification?
  - How difficult is it to collect this information in practice?
  - How much variation is there in each variable?
  - How much additional variation is there in each variable?

Confirming or rejecting your project goal

> Given data availability and data quality, is the project still feasible?

> Given the descriptions of your outcome variable, is the problem prevalent or pressing enough?

Further refining the outcome Y

> What choices need to be made in case the outcome Y needs to be defined based on the data?
  - What counts as early dropout?
  - What should be considered a treatment success?
  - What cut-off to choose when defining a satisfied vs unsatisfied customer?

> How do different definitions of Y impact the balance in the outcome?

> Has the user been carefully consulted in defining Y?

Further scoping your project

> Are there outlying X-values which are better left out-of-scope?

> Is the problem perhaps more relevant in a subpopulation?

> Has the user been carefully consulted in scoping the project?

Input for data preparation and cleaning

> Which variables to exclude as they would not be available at the moment of prediction or classification?
> Where and how to impute missing values?
> How to correct infeasible (combinations of) values?
> Which variables to exclude for having little to no variation?
> Which variables to exclude for being highly collinear with other variables?
> Which variables to exclude for containing information that would be too difficult to collect in practice?

Feature engineering and model considerations

> Which input variables do and do not have an individual relation with Y?

> In what (possibly non-linear) way are input variables related to Y?

> What combination of input variables are related to Y?

> Which input variables seem to be most strongly related to Y?
  - when is this information available?
  - what does this mean for when the model can be applied?