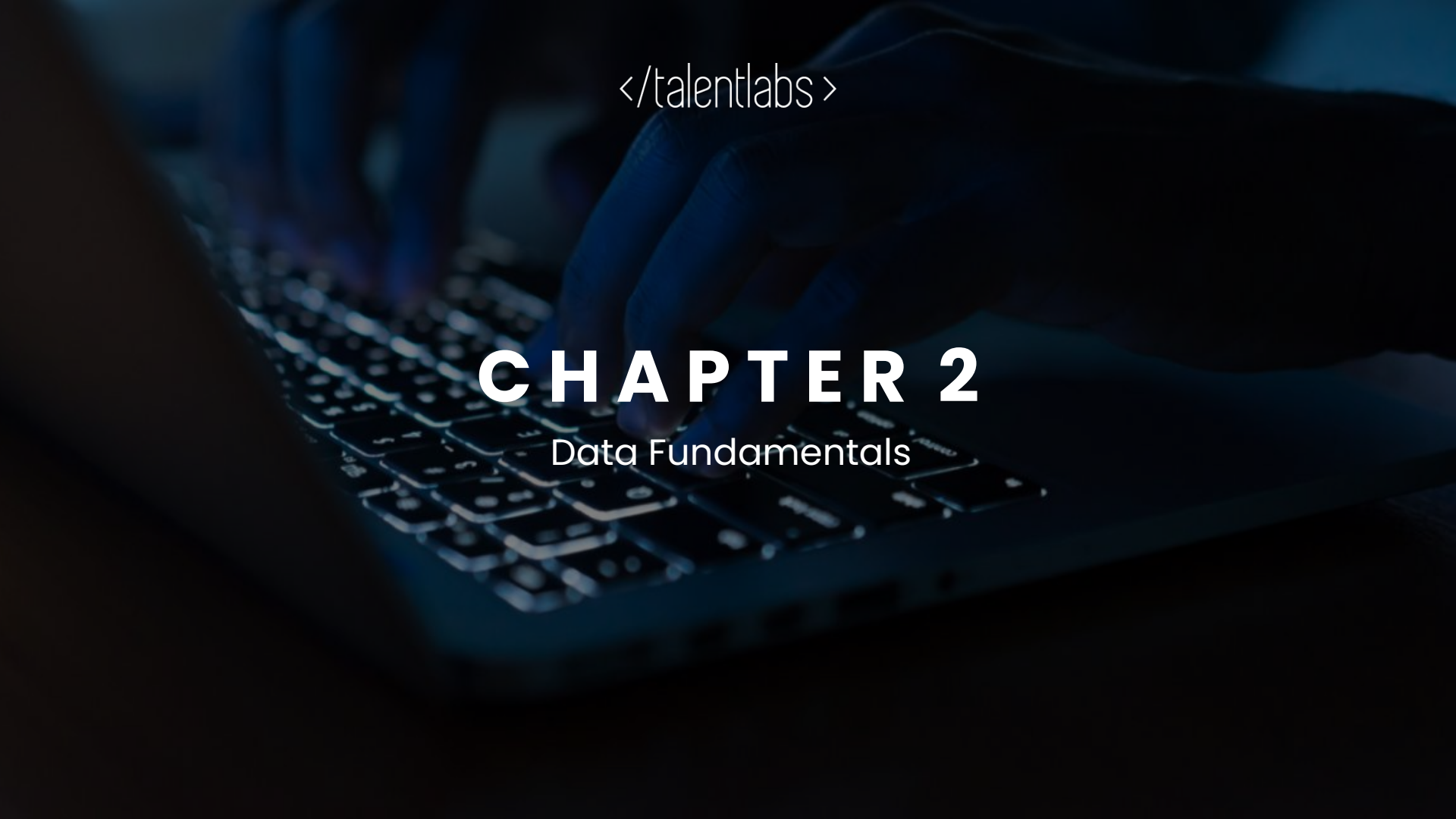

</talentlabs>
CHAPTER 2
Data Fundamentals

Learning Objectives

- Appreciate how data is used in everyday life
- Explain the difference between qualitative and quantitative data
- Understand the difference between big and small data
- Define metadata

Agenda

- Data in everyday life
- Data formats
- Data types
- Big vs small data
- How data is stored

</talentlabs>

Data in everyday life

What is data?

Data in everyday life

There were 5 exabytes of information created between the dawn of civilization through 2003, but that much information is now created every two days.

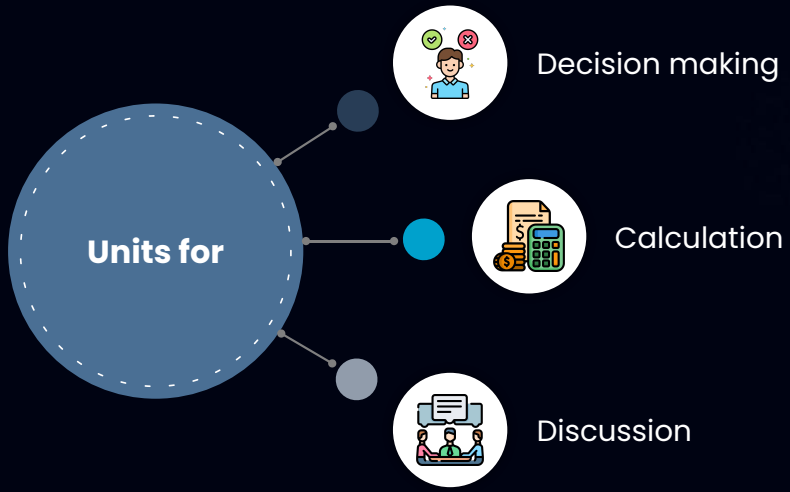
– Eric Schmidt,
Executive Chairman at Google

1 exabyte
=
1,000,000,000
gigabytes

</talentlabs>

What is data?

Collection of *facts* or *information*

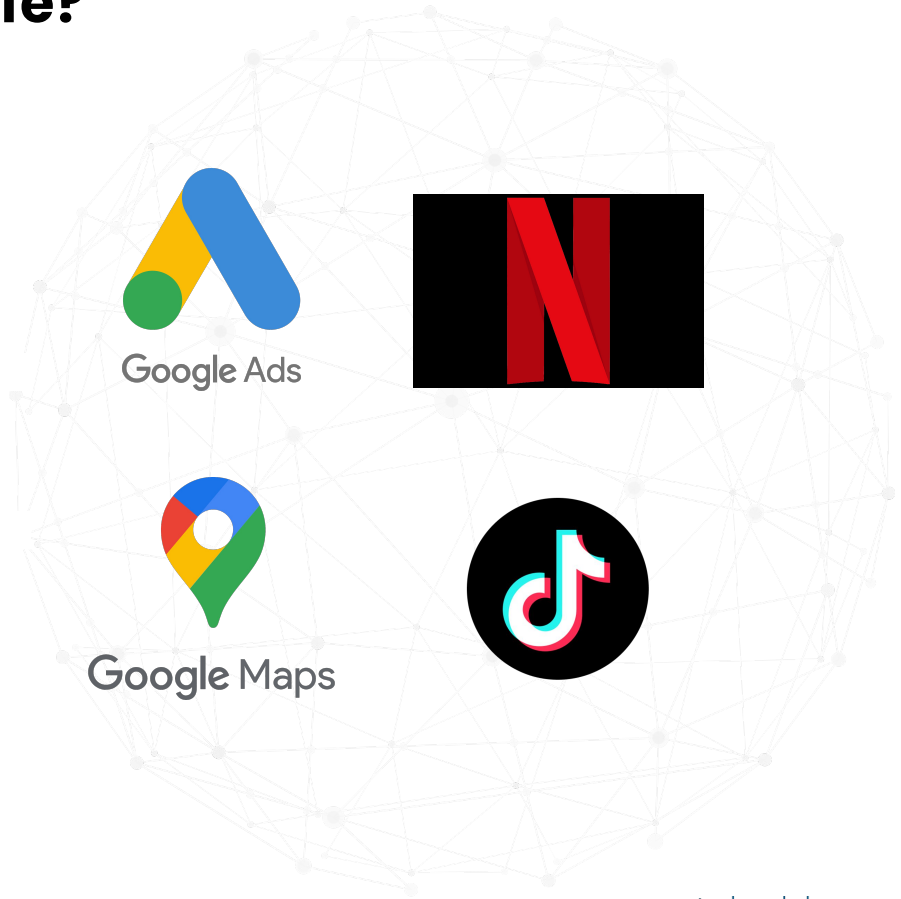


How do you use data everyday?

</talentlabs>

How does data affect your daily life?

Entertainment	Movie recommendations
Advertisement	Cookies + personal ads
Traffic data	Optimal route to get from A to B
Social media	Personalised content



</talentlabs>

There were 5 exabytes of information created between the dawn of civilization through 2003, but that much information is now created every two days.

– Eric Schmidt,
Executive Chairman at Google

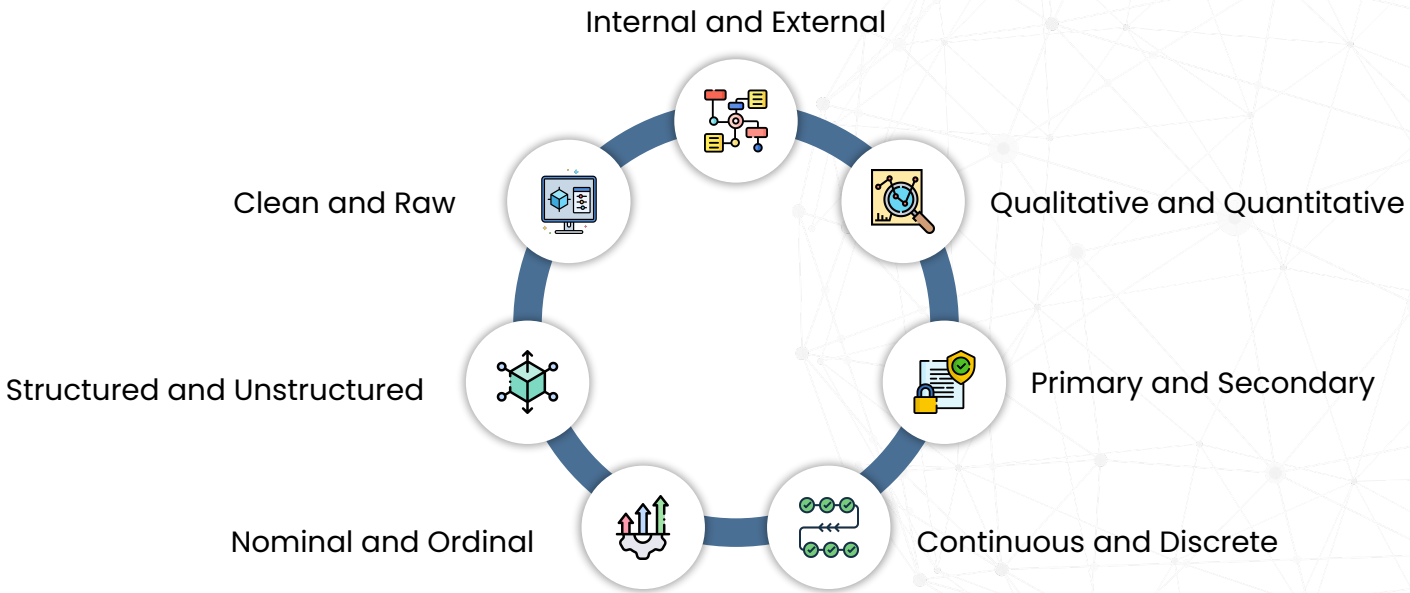
Data opens up a lot of opportunities

</talentlabs>

Data formats

- Qualitative vs quantitative data
- Structured vs unstructured data

Data Formats



</talentlabs>

Qualitative vs Quantitative

Qualitative

Descriptive data (size, colour, quality) –
analysed through categorisation and interpretation

Quantitative

Numeric data –
analysed through statistical methods

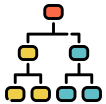
</talentlabs>

Qualitative vs Quantitative

Manufacturer	Model	Vehicle type	Sales in thousands	Engine size	Horsepower
Ford	F-Series	Car	541	4.6	220
Ford	Explorer	Car	277	4.0	210
Toyota	Camry	Passenger	248	2.2	133
Ford	Taurus	Passenger	246	3.0	155
Honda	Accord	Passenger	231	2.3	135
Dodge	Ram Pickup	Car	227	5.2	230
Ford	Ranger	Car	221	2.5	119
Honda	Civic	Passenger	200	1.6	106
Dodge	Caravan	Car	182	2.4	150
Ford	Focus	Passenger	176	2.0	107

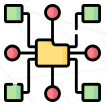
</talentlabs>

Internal vs External



Internal

Data existing within the organisation



External

Data existing outside the organisation

</talentlabs>

Primary vs Secondary

Primary

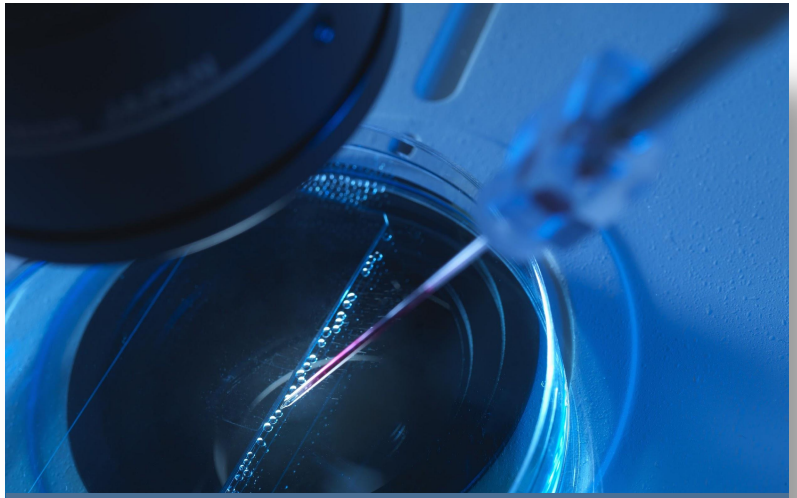
Gathered by the researcher first-hand

Secondary

Gathered by other researchers

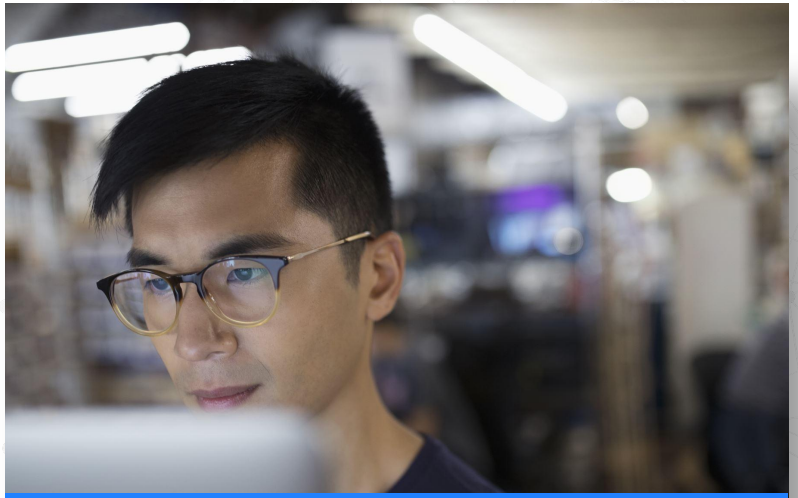
</talentlabs>

Primary vs Secondary



Primary

Collecting data from an experiment first-hand



Secondary

Downloading a dataset online collected by someone else

</talentlabs>

Continuous vs Discrete

Continuous

Can hold any value
(profit, time)

Discrete

Limited number of values
(number of languages spoken by an employee, number of students in a class)

</talentlabs>

Nominal vs Ordinal

Nominal

Qualitative data that isn't categorised in a particular order (colour, brand, material)

Ordinal

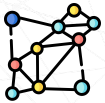
Qualitative data with a set order (quality – ranked from poor to excellent, education level)

</talentlabs>

Clean vs Raw

Clean

Complete, consistent and accurate data that is ready for analysis



Raw

(Or dirty data) Not cleaned and unprocessed – may contain errors, poor data entry, etc

</talentlabs>

Clean vs Raw

Raw		
Spotted Shark Species	**Country**	**Date**
Great White	USA	11/04/22
White Shark	Australia	28/02/21
Tiger	South Africa	03/28/22
Bull	Australia	12/01/2022
White	America	23/12/21
Tiger Shark	australia	11/11/21

Clean		
Spotted Shark Species	**Country**	**Date**
Great White	USA	11/04/2022
Tiger	South Africa	28/03/2022
Bull	Australia	12/01/2022
Great White	USA	23/12/2021
Tiger	Australia	11/11/2021
Great White	Australia	28/02/2021

</talentlabs>

Structured vs Unstructured

Structured

Organised in a format – rows and columns
(e.g. excel tables and csv files)

Unstructured

Not organised in a predefined format
(e.g. video and audio files)

</talentlabs>

Structured vs Unstructured

Structured		
Spotted Shark Species	**Country**	**Date**
Great White	USA	11/04/2022
Tiger	South Africa	28/03/2022
Bull	Australia	12/01/2022
Great White	USA	23/12/2021
Tiger	Australia	11/11/2021
Great White	Australia	28/02/2021

Structured vs Unstructured

Structured

Easy to organise

Easy to search

Easy to analyse

Often quantitative data

Unstructured

Unorganised

Difficult to search

Challenging to analyse

Often qualitative data

</talentlabs>

Challenges of using unstructured data

Lack of structure = difficult to work with

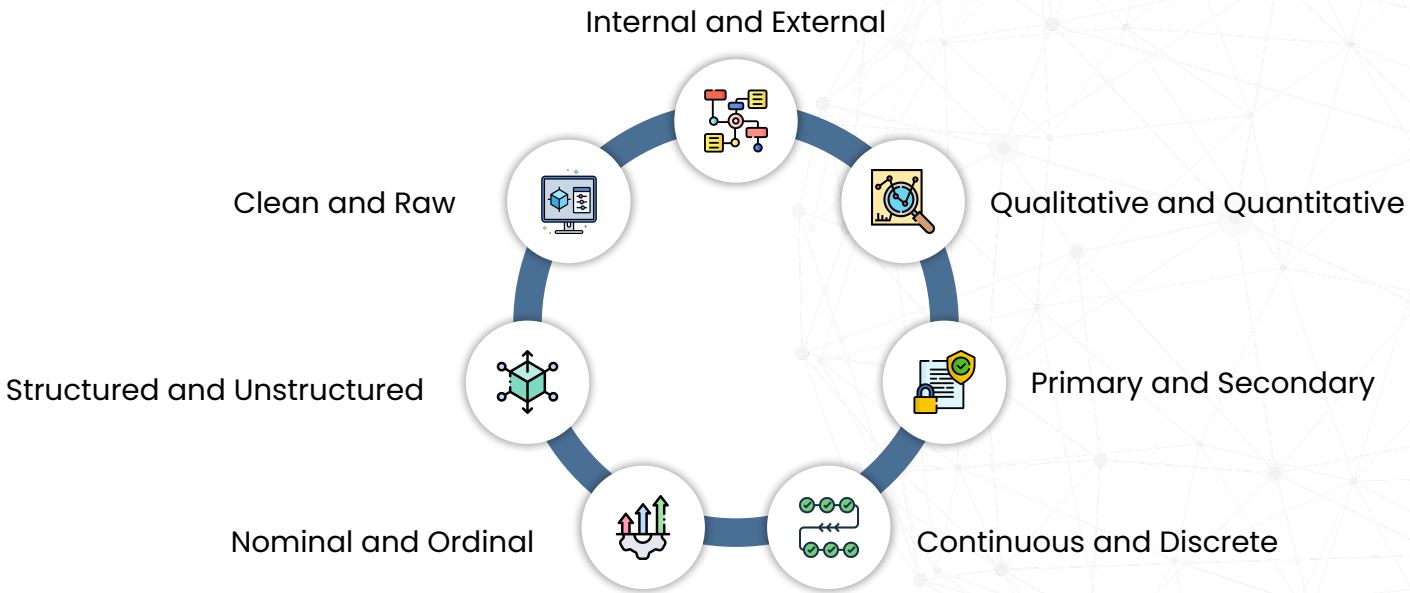
Machine learning research progress is changing this

Data bias

Skewed and unreliable results

</talentlabs>

Data Formats



</talentlabs>

Data types

Core data types

Other common data types

Core data types

Data type	Description	Example
Integer	Positive or negative whole numbers	8
Float	Numbers with a decimal place	8.674
Character	Single character letter, number or symbol	%
String	Text	‘Hello world’
Boolean	True or False (1 or 0)	False

Other popular data types

- Timestamp
- Date
- Datetime
- Geographic

</talentlabs>

Other popular data types

Timestamp	Datetime	Date	Geographic
1970-01-01 00:00:01 UTC	2021-12-31 23:59:59	2021-12-31	(47.65100, -122.34900, 4326)

Timestamp vs datetime

- Timestamp – has a shorter range (~1970-2038)
- Timestamp has time zone parameter

Geographic data is typically

- Point
- LineString
- Polygon

</talentlabs>

Big vs small data

Difference between small and big data

Challenges and benefits of big data

Small and Big Data

Small

Specific, simple enough for conventional analysis techniques, accessible and concise (<100s of gigabytes)

Big

Less specific, large chunks of structured and unstructured data (terabytes<)

Differ in

Collection, volume, analysis, quality, structure, storage, security, processing

</talentlabs>

Small and Big Data Examples

Small

Football scores

Company's financial reports

Sales data

Big

Stock exchange

Healthcare

Recommendation engines

</talentlabs>

Big Data Three Vs

Volume

Size of the data –

how big the data is in storage size

Variety

Variety in data sources –

the number of sources and data types

Velocity

Growth of data –

how fast data is coming in

</talentlabs>

Challenges and benefits of big data

Challenges

Irrelevant information

Accessibility

Insights hidden (needle in a haystack)

Security

Benefits

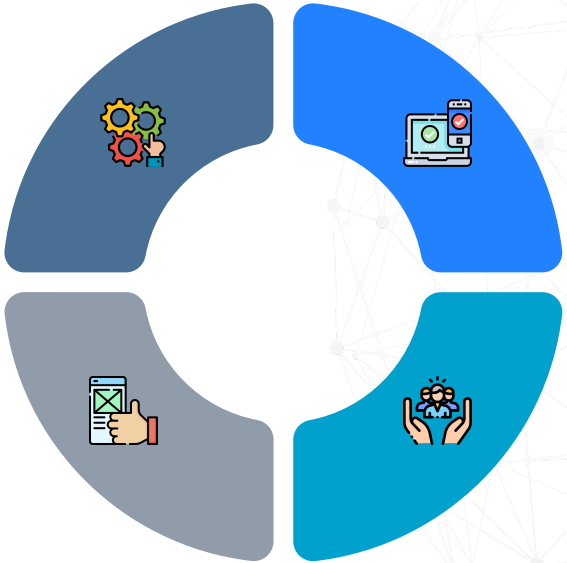
When challenges overcome can yield tremendous results

</talentlabs>

Big Data Solutions – What to look for

Automation

Automated tools vs manual data collection



Accessibility

Usability for different levels of expertise, good user experience design

Usability

Easy interpretability, quicker insights

Adoptability

Fast to learn and adopt

</talentlabs>

How data is stored

- Data file formats
- Data storage
- Metadata and data schema

Data file formats

Depend on type of data

Examples:

- Tabular (CSV, XLSX)
- Non-tabular (TXT)
- Image (PNG, JPEG)

Proprietary files

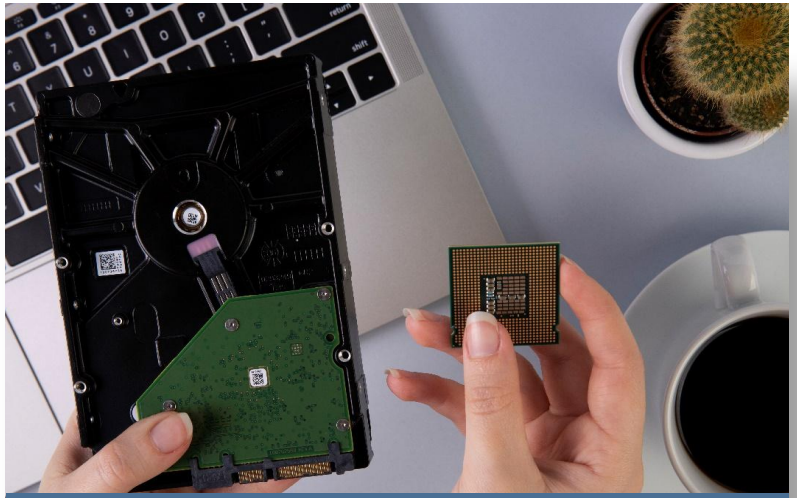
Made to be opened by specific software

Best practice

Use most accessible file formats

</talentlabs>

Data storage



Direct

SSD, HDD

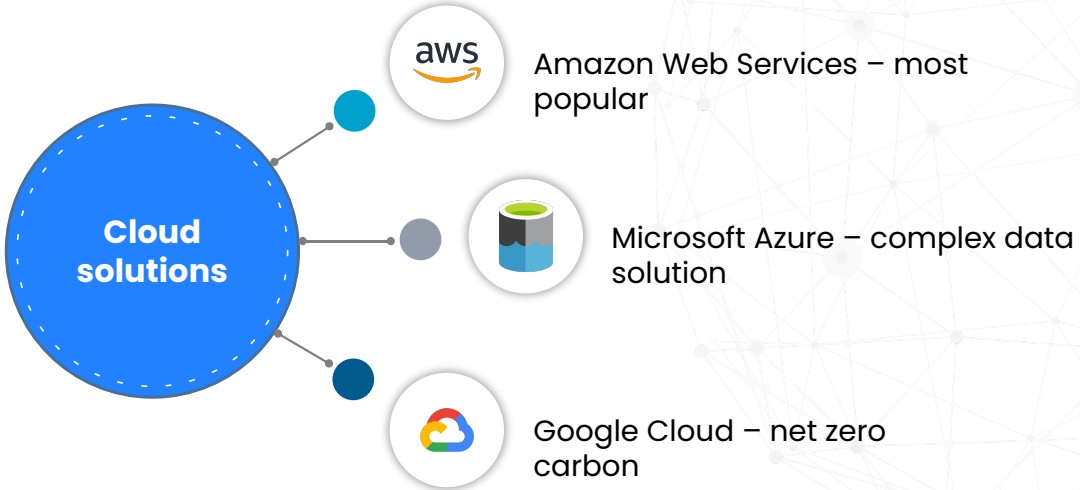
Network-based

Cloud

</talentlabs>

Popular Cloud Solutions

A *data lake* is a central repository that stores and protects large volumes of data – structured and unstructured



</talentlabs>

Metadata and Data Schema

Metadata

Metadata is **data about data** – information about stored data that helps analysts understand the contents of the data files

Metadata is useful **irrespective of the dataset size.**

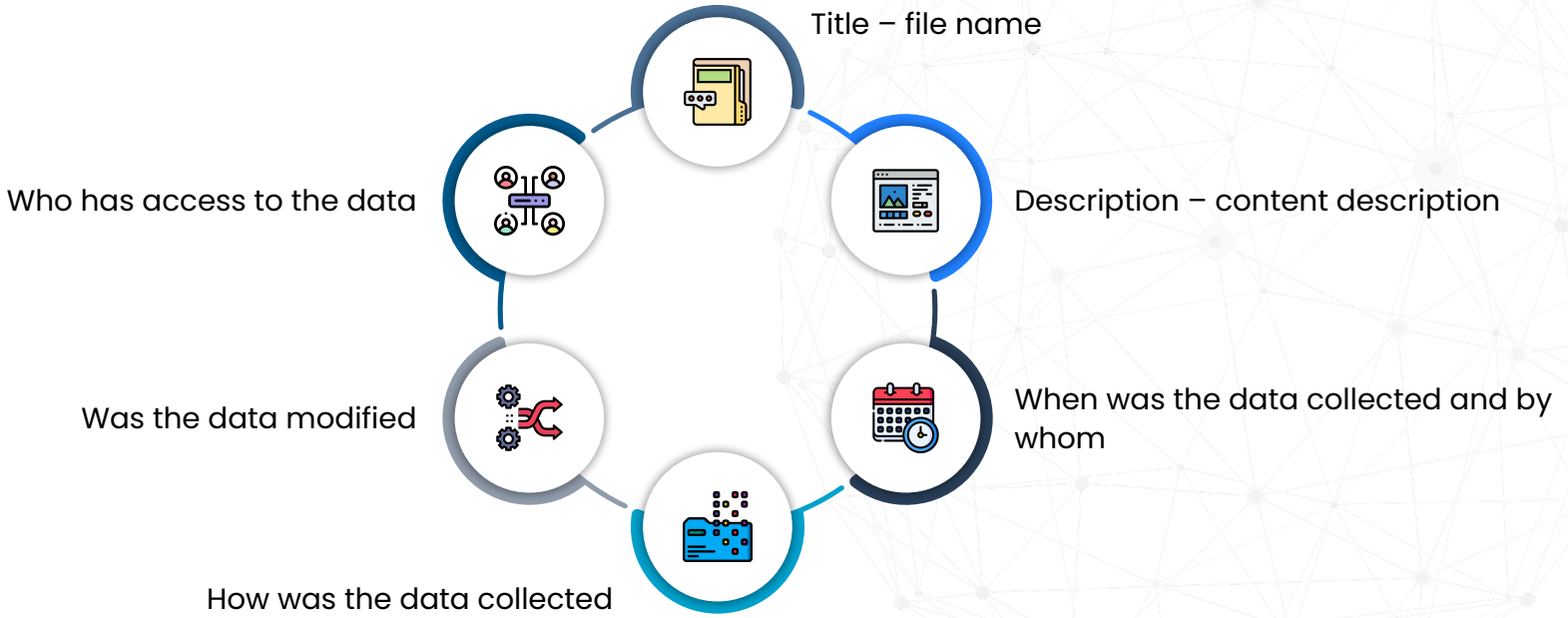
Answers the what, when, why, who, where and how of the data

Data schema

Data schema defines the **organisation** of data and the **relationship** between different tables within a database.

Commonly use visual representations to display organisation of data.

What information is included in metadata?



</talentlabs>

Example Metadata

Manufacturer	Model	Vehicle type	Sales in thousands	Engine size	Horsepower
Ford	F-Series	Car	541	4.6	220
Ford	Explorer	Car	277	4.0	210
Toyota	Camry	Passenger	248	2.2	133

Dataset containing sales information of different cars. Specific information about the cars, such as the model type are included.

The data was collected by Talent Labs on the 10/06/2022.

The data was collected via web scraping using Python and Selenium libraries.

Column	Data Type	Description
Manufacturer	String – categorical	Car manufacturer
Model	String – categorical	Model of the car
Vehicle type	String – categorical	Type of vehicle (either car or passenger)
Sales in thousands	Integer	Worldwide sales in thousands (USD)
Engine size	Float	Engine size (L)
Horsepower	Integer	Horsepower

Other metadata examples

Photos

Filename, date, time, camera and location

Emails

Sender, date and time sent and hidden data

Books

Title, author's name, contents, publisher details

</talentlabs>

Why make a metadata file?

Provides **context**

Increase data reliability

Maintains **consistency**

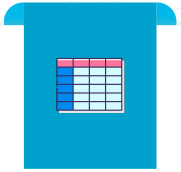
Helps new users quickly understand the information

Improves third party **trust**

</talentlabs>

What information is included in a data schema?

All relevant data



Column names and data types for each table

Database object unique keys

Consistent formatting for data entry

</talentlabs>

Database schema types

Two main types

Logical

How data is organised in tables and how columns link together

Physical

Represents how data is stored

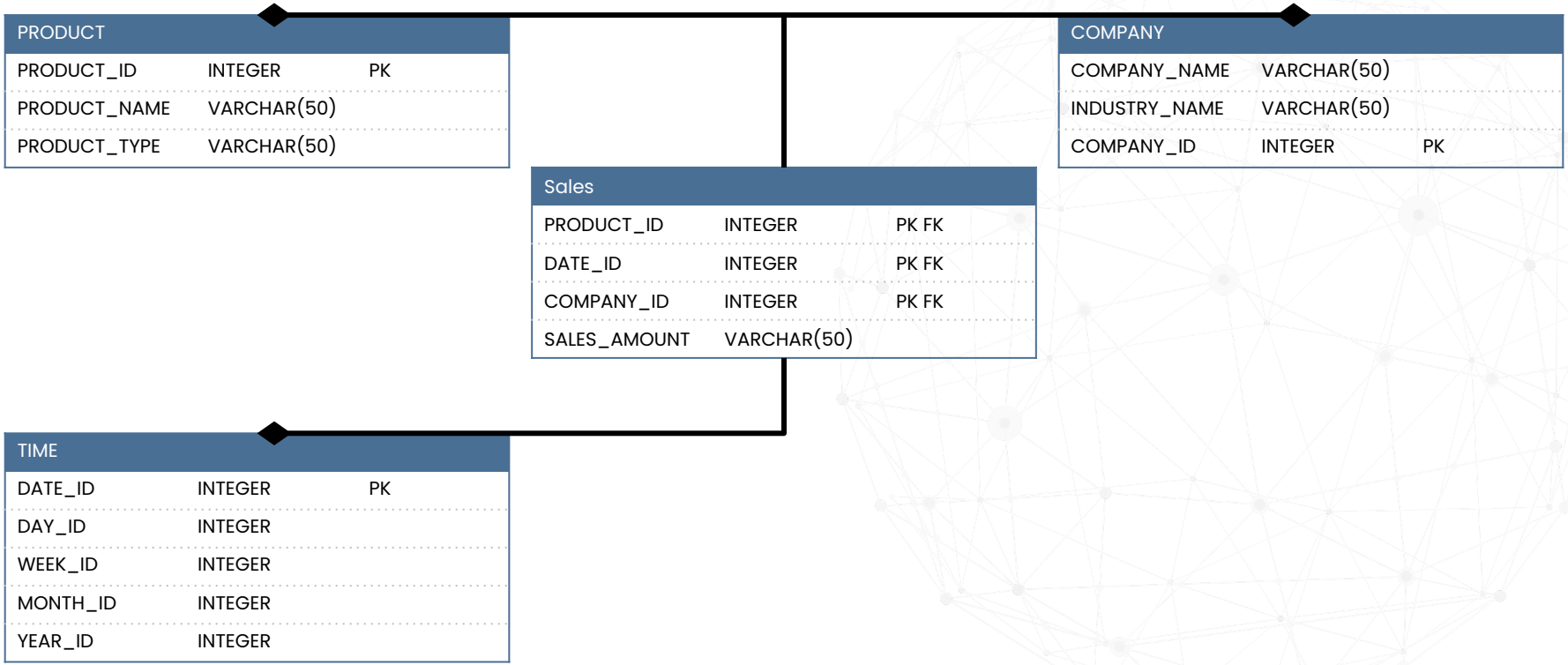
Primary key

Identify a record in the table

Foreign key

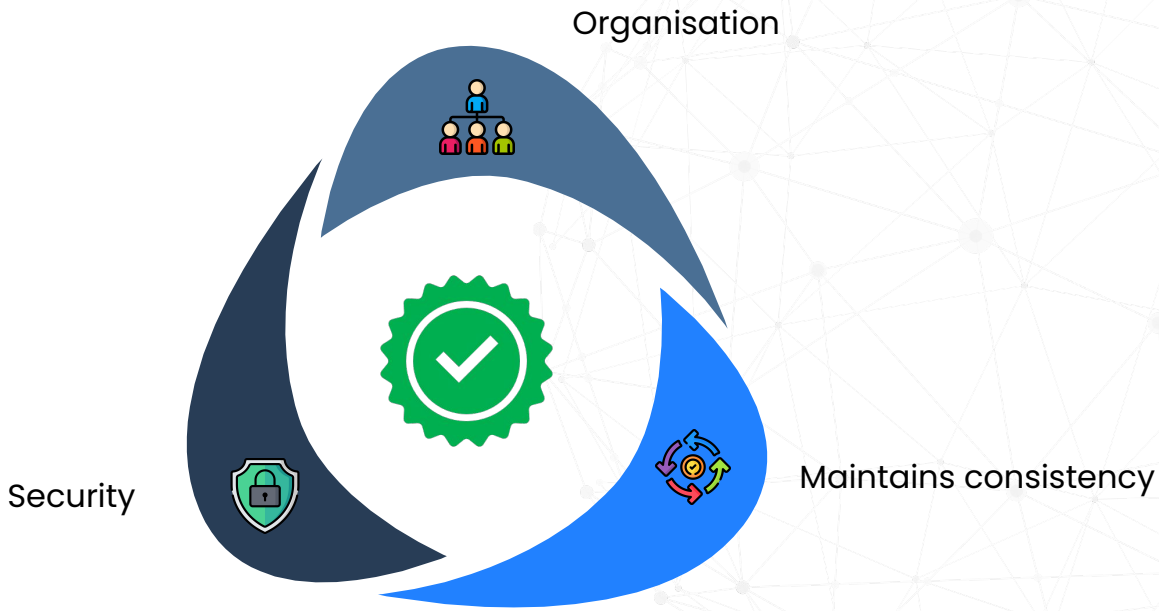
Primary key for another table

Reading Database Schemas



</talentlabs>

Why use data schemas?

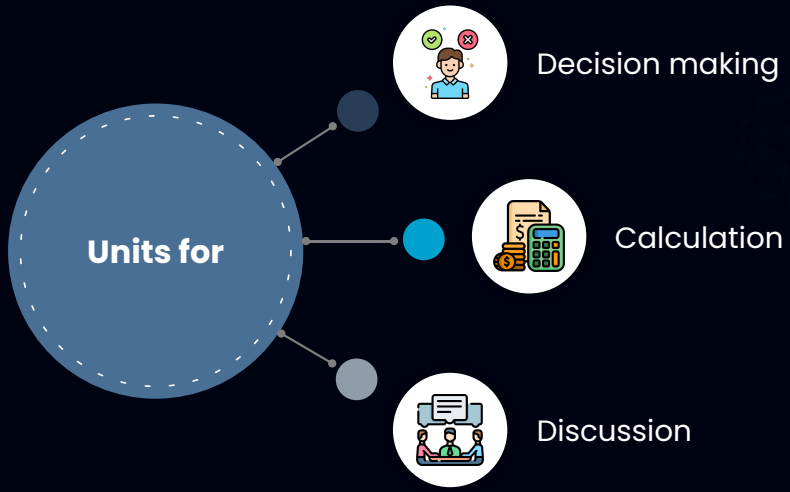


</talentlabs>

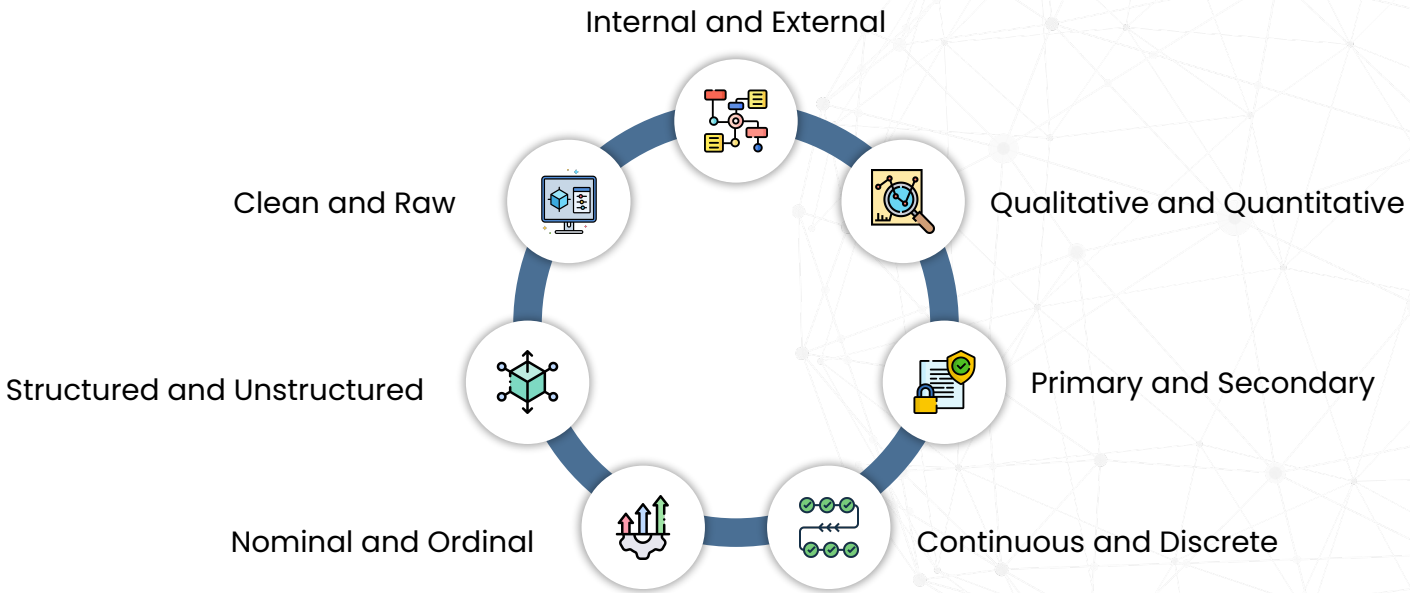
Summary and Assignment

What is data?

Collection of *facts* or *information*



Data Formats



</talentlabs>

Qualitative vs Quantitative

Manufacturer	Model	Vehicle type	Sales in thousands	Engine size	Horsepower
Ford	F-Series	Car	541	4.6	220
Ford	Explorer	Car	277	4.0	210
Toyota	Camry	Passenger	248	2.2	133
Ford	Taurus	Passenger	246	3.0	155
Honda	Accord	Passenger	231	2.3	135
Dodge	Ram Pickup	Car	227	5.2	230
Ford	Ranger	Car	221	2.5	119
Honda	Civic	Passenger	200	1.6	106
Dodge	Caravan	Car	182	2.4	150
Ford	Focus	Passenger	176	2.0	107

Structured vs Unstructured

Structured		
Spotted Shark Species	**Country**	**Date**
Great White	USA	11/04/2022
Tiger	South Africa	28/03/2022
Bull	Australia	12/01/2022
Great White	USA	23/12/2021
Tiger	Australia	11/11/2021
Great White	Australia	28/02/2021

Core data types

Data type	Description	Example
Integer	Positive or negative whole numbers	8
Float	Numbers with a decimal place	8.674
Character	Single character letter, number or symbol	%
String	Text	‘Hello world’
Boolean	True or False (1 or 0)	False

Small and Big Data

Small

Specific, simple enough for conventional analysis techniques, accessible and concise (<100s of gigabytes)

Big

Less specific, large chunks of structured and unstructured data (terabytes<)

Differ in

Collection, volume, analysis, quality, structure, storage, security, processing

</talentlabs>

Big Data Three Vs

Volume

Size of the data –

how big the data is in storage size

Variety

Variety in data sources –

the number of sources and data types

Velocity

Growth of data –

how fast data is coming in

</talentlabs>

Metadata and Data Schema

Metadata

Metadata is **data about data** – information about stored data that helps analysts understand the contents of the data files

Metadata is useful **irrespective of the dataset size.**

Answers the what, when, why, who, where and how of the data

Data schema

Data schema defines the **organisation** of data and the **relationship** between different tables within a database.

Commonly use visual representations to display organisation of data.

Assignment Information

Multiple choice questions

Hands-on assignment

</talentlabs>