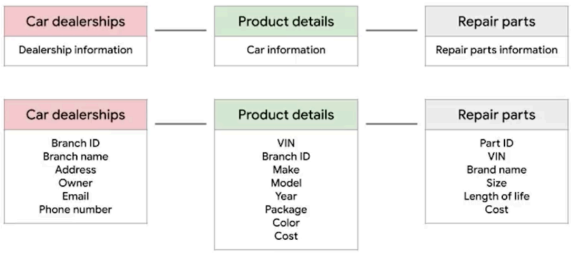
Database Features & Components

Freitag, 11. Juli 2025 10:29

Metadata
L> data about data
L> a data analyst thinking about how he's analysing data

Basic Database Structure Example
+ this example is called a relational database



Relational Database
L> a database that contains a series of related tables that can be connected via their relationships
L> one of the same fields must exist inside both tables
L> in our example it's the field *"Branch ID"*

"In a non-relational table, you will find all of the possible variables you might be interested in analysing all grouped together. This can make it really hard to sort through. Relational databases simplify a lot the analysis process & make it easier to find data & use it across an entire database."

Normalisation
L> the process of organising data in a relational database
L> eliminates data redundancy
L> increases data integrity
L> reduces complexity in a database

Primary Key
L> identifier that references a column in which each value is unique
L> think of it as a unique identifier for each row in the table
L> in our example *"Branch ID"* is the primary key
L> for the *"Product Details* Table*"*, *"VIN"* is our primary key
L> if we use primary keys, they should be unique
L> there can only be one primary key

Key Features
L> used to ensure data in a specific column is unique
L> uniquely identifies a record in a relational database table
L> only one primary key is allowed in a table

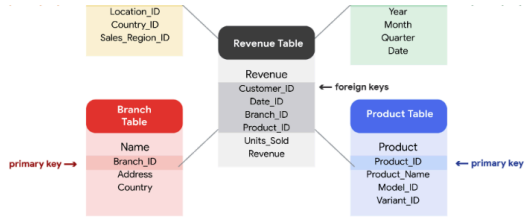
Foreign key
L> a field within a table that is a primary key in another table
L> how one table can be connected to another table
L> in our example it's *"Part ID"*
L> each row in our *"Repair Parts Table"* represents one unique part
L> all the other keys are foreign keys, allowing other tables to connect
L> there can be multiple foreign keys

Key Features
L> a column or a group of columns in a relational database, that provides a link between the data in two tables
L> refers to the field in a table that's the primary key of another table
L> more than one foreign key is allowed to exist in a table

Composite Key
L> a primary key constructed using multiple columns in a table

Some tables **don't require** a primary key - like a revenue table

Datei a



SQL = Game-Changer, when working with multiple databases

Metadata Management

Freitag, 11. Juli 2025 10:29

Metadata
L> data about the data
L> used in database management
L> helps data analysts interpret contents of data within database

First Example
L> check the extra-information inside a photo from our gallery
L> every time we send a message - and check the message details
L> it might include time, location, and more

Spreadsheets also contain all kinds of Metadata
L> file type, file size, owner, last modified date, download permissions

Metadata Repositories
L> specialised databases created to store and manage Metadata
L> to bring together multiple sources for data analysis
L> especially important, when using third-party data

Data Governance
L> a process to ensure the formal management of a company's data assets

3 Types of Metadata

Descriptive
L> describes a piece of data
L> can be used to identify it at a later point of time
L> e.g. the code on a spine of a book, called ISBN

Structural
L> indicates how a piece of data is organised
L> whether it is part of one, or more than one, data collection
L> e.g. how the pages of the book are put together into chapters
L> also keeps track of the relationship between two things

Administrative
L> indicates the technical source of a digital asset
L> e.g. the picture - file-type, file-size

The benefits of Metadata

Reliability
L> helps us to confirm the reliability of the data we use
L> make sure it's accurate, precise, relevant, timely

Consistency
L> having survey data for two different sources
L> using metadata to make sure, same collections methods were applied
L> make sure it's organised, classified, stored, accessed

Datei auswählen Keine ausgewählt

Access Data Sources

Freitag, 11. Juli 2025 10:29

Internal Data
L> often called primary data
External Data
L> often called secondary data

Comma Separated Values Files (.csv)
+ plain text files with organised table structure
+ including rows and columns
+ values in each row are separated by commas
+ widespread compatibility *(for import/export)*

We know how to do normal importing of .csv files.

Dynamic Import

IMPORTANGE function
L> enables range of cells in one spreadsheet to be duplicated in another
L> =IMPORTANGE ("URL", "sheet_name!cell_range")
L> allows to automatically update spreadsheets

IMPORTHTML Function
L> importing HTML tables or lists on a web page
L> this process is often called *"scraping"*

IMPORTDATA Function
L> sometimes data displayed on the web is in another form
L> form of comma- or tab-delimited file
L> besides this function is similar to IMPORTANGE function

Datei auswählen Keine ausgewählt

Public Data Sets

Freitag, 11. Juli 2025 10:29

Open data helps create a lot of public datasets.
You can access them to make data-driven decisions.
Here are some resources you can use to start:

Google Cloud Public Datasets
Allow data analysts access to high-demand public datasets, and make it easy to uncover insights in the cloud.

Dataset Search
Can help you find available datasets online with keyword searches.

Kaggle
Has an Open Data search function; help find datasets to practice with

Big Query
Hosts 150+ public datasets, access and use, public health datasets

Global Health Observatory data
You can search for datasets from this page or explore featured data collections from the World Health Organization.

The Cancer Imaging Archive (TCIA) dataset
Just like the earlier dataset, this data is hosted by the Google Cloud Public Datasets and can be uploaded to Big Query.

1000 Genomes
This is another dataset from the Google Cloud Public resources that can be uploaded to Big Query.

National Climatic Data Centre
The NCDC Quick Links page has a selection of datasets you can explore.

NOAA Public Dataset Gallery
The NOAA Public Dataset Gallery contains a searchable collection of public datasets.

UNICEF State of the World's Children
This dataset from UNICEF includes a collection of tables that can be downloaded.

CPS Labor Force Statistics
This page contains links to several available datasets that you can explore.

The Stanford Open Policing Project
This dataset can be downloaded as a .csv file for your own use.

Datei auswählen Keine ausgewählt

Sort & Filter Data

Freitag, 11. Juli 2025 10:29

Sorting Data
L> arranging data into a meaningful order
L> to make it easier to understand, analyse and visualise

Freezing a Row
L> select a row
L> go to the tab "view"
L> press on "freeze"
L> select "1 row"

Multiple Sorting
L> select whole sheet
L> go to the tab "data"
L> then "sort sheet"
L> then "advanced"
L> you can add several parameter
L> like state and city sorting together

Filtering
L> showing only the data that meets a specific criteria
L> while hiding the rest
L> go to "Data" and "Filter"
L> you can give a word from the sheet
L> and everything else will be hidden

Convert Data *(Text -> Numbers)*
L> select a column
L> go to "edit" - "find and replace"
L> remember to click on "Match entire cell contents"
L> you can type in a word and replace by a number
L> or a number by a text, probably more rare

Datei auswählen Keine ausgewählt

SQL - Big Query

Freitag, 11. Juli 2025 10:29

Big Query
+ a data warehouse
+ it belongs to Google
+ free version allows me to have 12 projects

+ you can access public data sets
+ you have a console to access *(query)* data

+ search for a data set in the **Explorer**
+ once you found something
L> use the three dots next to it
L> and press query

Basic Functions
SELECT *
FROM 'the_name_of_the_database'
WHERE column_name='word'

Name of the Database
Composed of..
Database_Name . Database_Table

If we wouldn't write SELECT *
L> we would have to write every field name

Be careful about the format
L> using ` or , or ; or >=

Good Example
```
SELECT
   duration,
   start_station_name
FROM
   `bigquery-public-data.london_bicycles.cycle_hire`
WHERE
   duration >= 1200;
```

We can also put function into SQL
L> like SUM()

Sometimes you need " " to avoid errors:
L> favorite_food = "Shepherd's pie"
L> because there is a ' in the text

Use -- to create single-line notes for yourself!
You can also use /* and */ for multi-line notes!
L> so that you remember what you were searching
L> later in excessive usage it will be important

You can use AS to give your columns new names!

Names should never have Spaces between them!
Use CamelCase or Snake_Case

Your Query shouldn't have more than 100 characters!
L> even through indentation is not a must *(using tabs & new lines)*
L> still use them to make it easier for others to read your code!
L> it's best practice!

Datei auswählen Keine ausgewählt

SQL - Regex

Freitag, 11. Juli 2025 10:29

Regex = Regular Expressions
Defined as flexible and precise pattern-matching tools

Cheat Sheet

https://jdhao.github.io/2019/02/28/sublime_text_regex_cheat_sheet/

Examples

```
-- Namen, die mit "A" beginnen
SELECT * FROM users
WHERE REGEXP_CONTAINS(name, r'^A');


-- Zeigt nur valide E-Mails (vereinfachtes Muster)
SELECT * FROM users
WHERE REGEXP_CONTAINS(email, r'^[\w._%+-]+@[A-Za-z0-9.-]+\.[A-Za-z]{2,}$');


-- Nur Telefonnummern mit exakt 10 Ziffern
SELECT * FROM contacts
WHERE REGEXP_CONTAINS(phone, r'^\d{10}$');


-- Sonderzeichen aus Telefonnummern entfernen
SELECT REGEXP_REPLACE(phone, r'[^0-9]', '') AS cleaned_phone
FROM contacts;

-- PLZ (Postleitzahlen) mit 5 Ziffern extrahieren
SELECT REGEXP_EXTRACT(address, r'\b\d{5}\b') AS zip_code
FROM customers;


-- Zeilen finden, die eine Zahl enthalten
SELECT * FROM products
WHERE REGEXP_CONTAINS(description, r'\d');


-- Nur URLs mit "https"
SELECT * FROM links
WHERE REGEXP_CONTAINS(url, r'^https://');


-- Alle Strings mit Bindestrich oder Unterstrich
SELECT * FROM strings
WHERE REGEXP_CONTAINS(col, r'[-_]');
```

Datei auswählen Keine ausgewählt

SQL - Create Dataset

Freitag, 11. Juli 2025 10:29

Actually it is not that hard
L> Create a Dataset
L> Dataset can be in existing project *(driven-tenure-465609-q2)*
L> Then go into the dataset
L> press "create table"
L> go for "upload"

Interesting is the point "Schema"
L> the CSV files don't have column names
L> information is like "Hannah, F, 29329"
L> name, gender, count, is missing

So into schema we enter following text:
name:string,
gender:string,
count:integer

New Function - ORDER BY
L> can make the columns be sorted
L> e.g. ORDER BY count DESC
L> means - order by count descending

New Function - LIMIT
L> can limit the number of appearing rows
L> e.g. LIMIT 5

Example with 2 new functions

```
SELECT
   *
FROM
   `driven-tenure-465609-q2.Babynames.names_2014`
WHERE
   gender ='M'
ORDER BY
   column DESC
LIMIT
   5
```

New Select-Option

```
SELECT
      AVG(tree_dbh)
```

Datei auswählen Keine ausgewählt

Glossary

Freitag, 11. Juli 2025 10:29

Terms and definitions for Course 3, Module 3

Administrative metadata: Metadata that indicates the technical source of a digital asset

CSV (comma-separated values) file: A delimited text file that uses a comma to separate values

Data governance: A process for ensuring the formal management of a company's data assets

Descriptive metadata: Metadata that describes a piece of data and can be used to identify it at a later point in time

Foreign key: A field within a database table that is a primary key in another table (Refer to primary key)

FROM: The section of a query that indicates where the selected data comes from

Geolocation: The geographical location of a person or device by means of digital information

Metadata: Data about data

Metadata repository: A database created to store metadata

Naming conventions: Consistent guidelines that describe the content, creation date, and version of a file in its name

Normalized database: A database in which only related data is stored in each table

Notebook: An interactive, editable programming environment for creating data reports and showcasing data skills

Primary key: An identifier in a database that references a column in which each value is unique (Refer to foreign key)

Redundancy: When the same piece of data is stored in two or more places

Schema: A way of describing how something, such as data, is organized

SELECT: The section of a query that indicates the subset of a dataset

Structural metadata: Metadata that indicates how a piece of data is organized and whether it is part of one or more than one data collection

WHERE: The section of a query that specifies criteria that the requested data must meet

World Health Organization: An organization whose primary role is to direct and coordinate international health within the United Nations system

Datei auswählen Keine ausgewählt