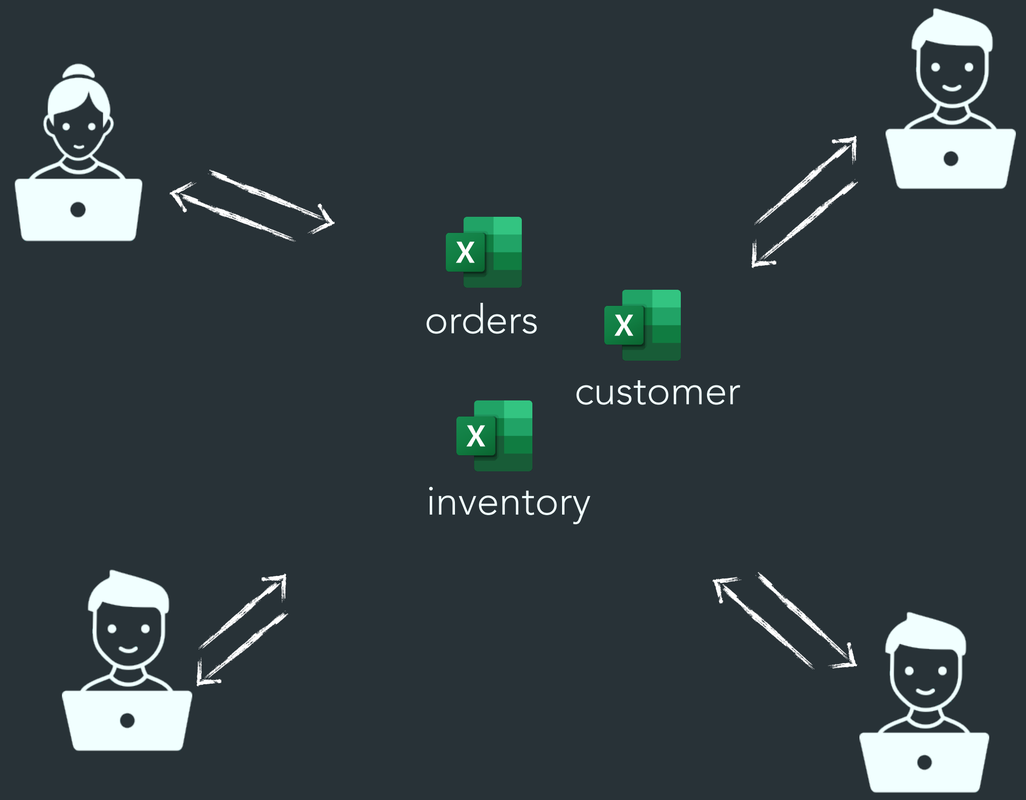
kokchun giang

turning data into valuable insights using **SQL**

an ice cream startup, Swedish Glassiker,
is using **excel sheets** to store data

okay in the beginning, but
the company grew …

each excel file is **shared** across the team



common **problems** that arose when sharing excel sheets in this way

data duplication

accidentally creating duplicates

manually updating in several places

inconsistent data

relationships

hard to manage manually between customer and order

team manually links orders to right customer

performance

performance issues when excel files grow larger

many more problems …

example of **inconsistent data** due to manual input

ID	Name	Email	Phone	Address
1	Alice Frost	alice@example.com	555-1234	123 Main St
2	Bob Cone	bob@example.com	555-5678	456 Maple Ave
3	Alice F.	alice_f@example.com	555-1234	123 Main St
4	Carla Scoop	carla@example.com	555-8765	789 Oak Blvd

is this same person?

now which Alice is linked to
what orders in the orders table?

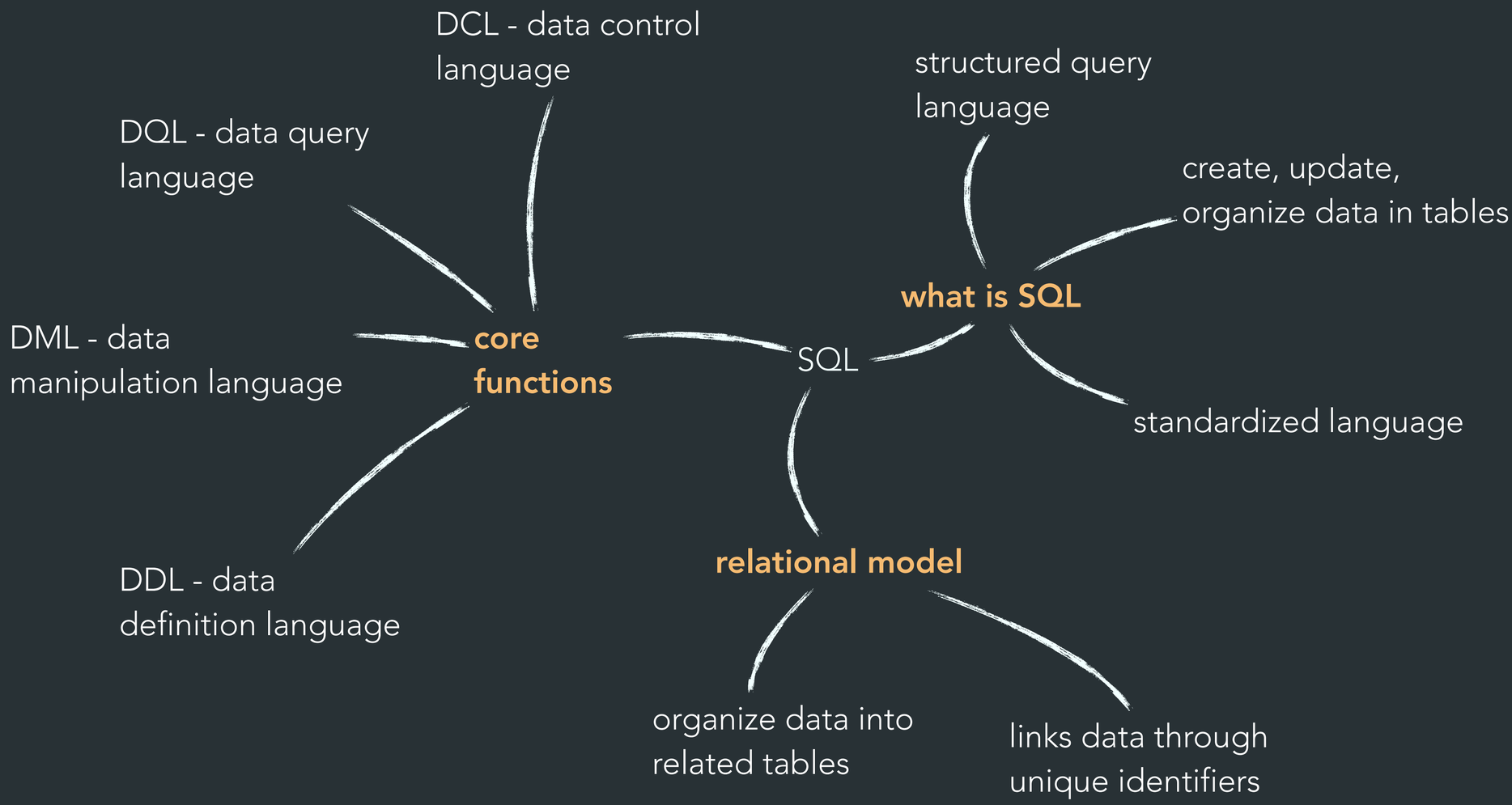
using **relational databases & SQL** we can handle many of these issues

defined relationships in SQL tables → ensure data consistency

optimized for large volumes of data → scalable and efficient querying of data

data constraints like data types and unique values → automatic validation reduces error

• • •


DCL - data control language
structured query language
DQL - data query language
create, update, organize data in tables
what is SQL
DML - data manipulation language
core functions
SQL
standardized language
DDL - data definition language
relational model
organize data into related tables
links data through unique identifiers

meet **duckdb**, a modern powerful database management system for analytics

OLAP
optimized for intensive analytical queries

high performance
on your own machine, can handle large datasets

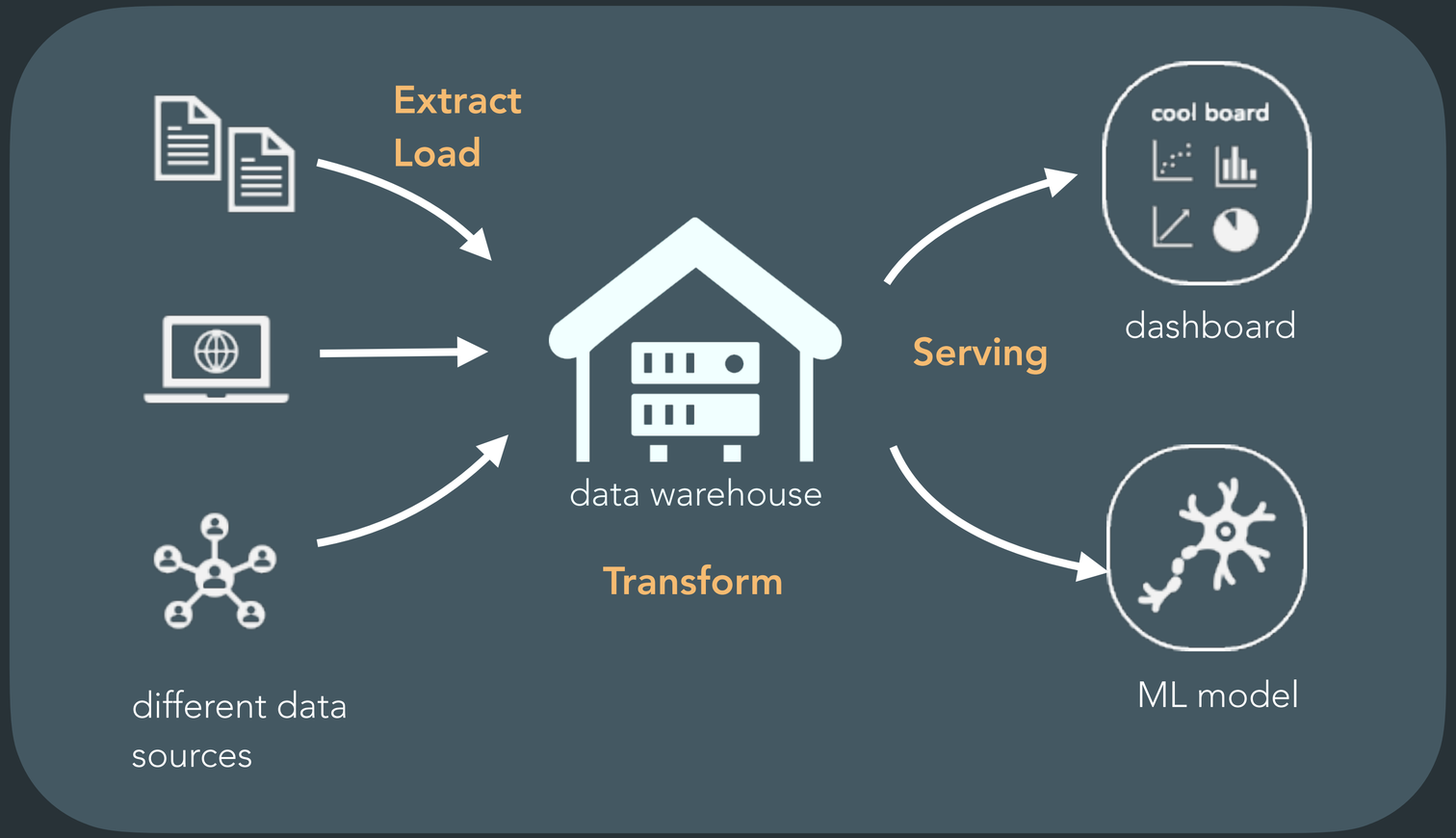
embedded database
no need for separate server, database contained in a file

integrates
with other tools in the data science ecosystem such as python, pandas and data frames

data analysis
can run complex queries for analytics and reporting

data transformations
in an ETL pipeline to serve BI and AI

a data engineering pipeline with an OLAP database as a **data warehouse**



duckdb could work as a lightweight data warehouse for small to medium-sized data