Introduction to Databases

Storing Massive Amounts of Data Safely + Efficiently

Why can’t we just use Excel files?

Why can’t we just use Excel files?

Limits to how much data we can store in memory.

What's wrong with just storing csv files?

What's wrong with just storing csv files?

Require cleaning (data types)

Not easily updated in real time

Similar memory issues during loading (for large sizes)

In industry, most of the time we are making analyses by pulling data from a dynamic resource

databases

What do databases do?

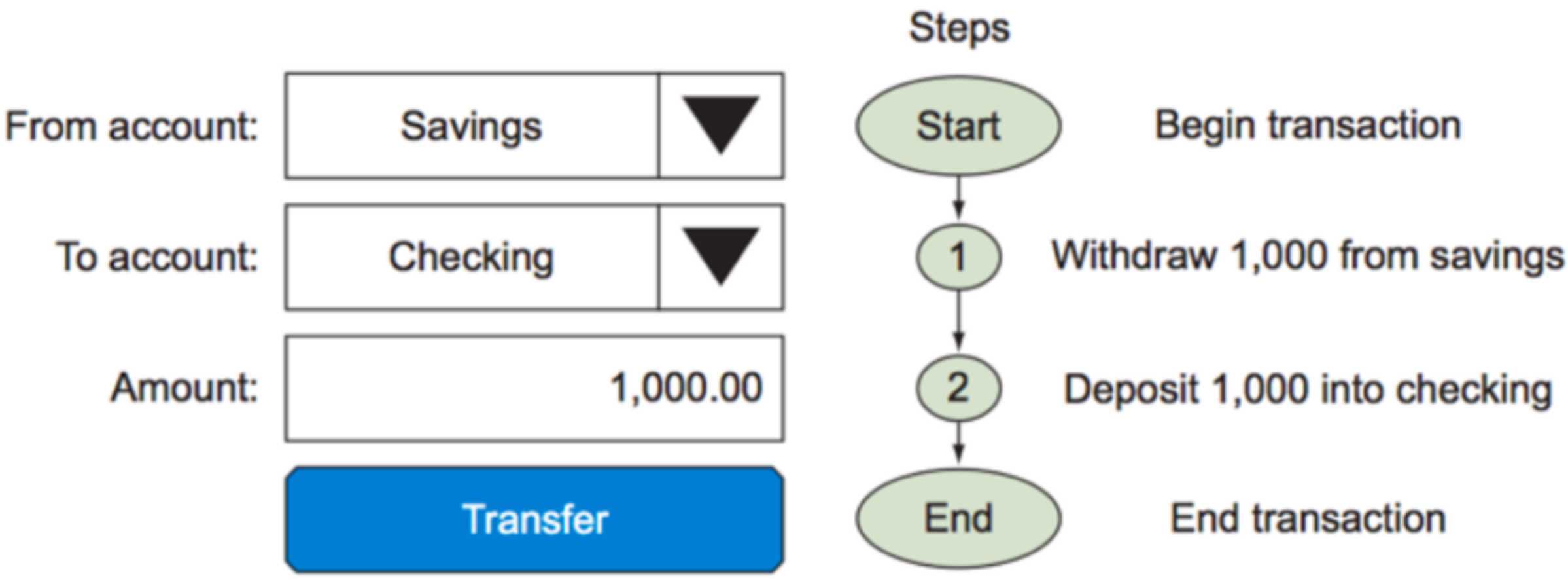
Manage storing and querying data

Retrieve data using SQL (Structured Query Language)

Create rules and structures to keep data clean

Data types must be well-defined — no Pandas-type cleaning

Transactional Integrity



Any change in a database = transaction

All transactions must be robust

Types of Databases

Relational

NoSQL

Key:Value Stores

Relational Databases

Data is organized in tables

Similar to Pandas dataframes

Each table has a primary key

Unique value for that row, specific to that table

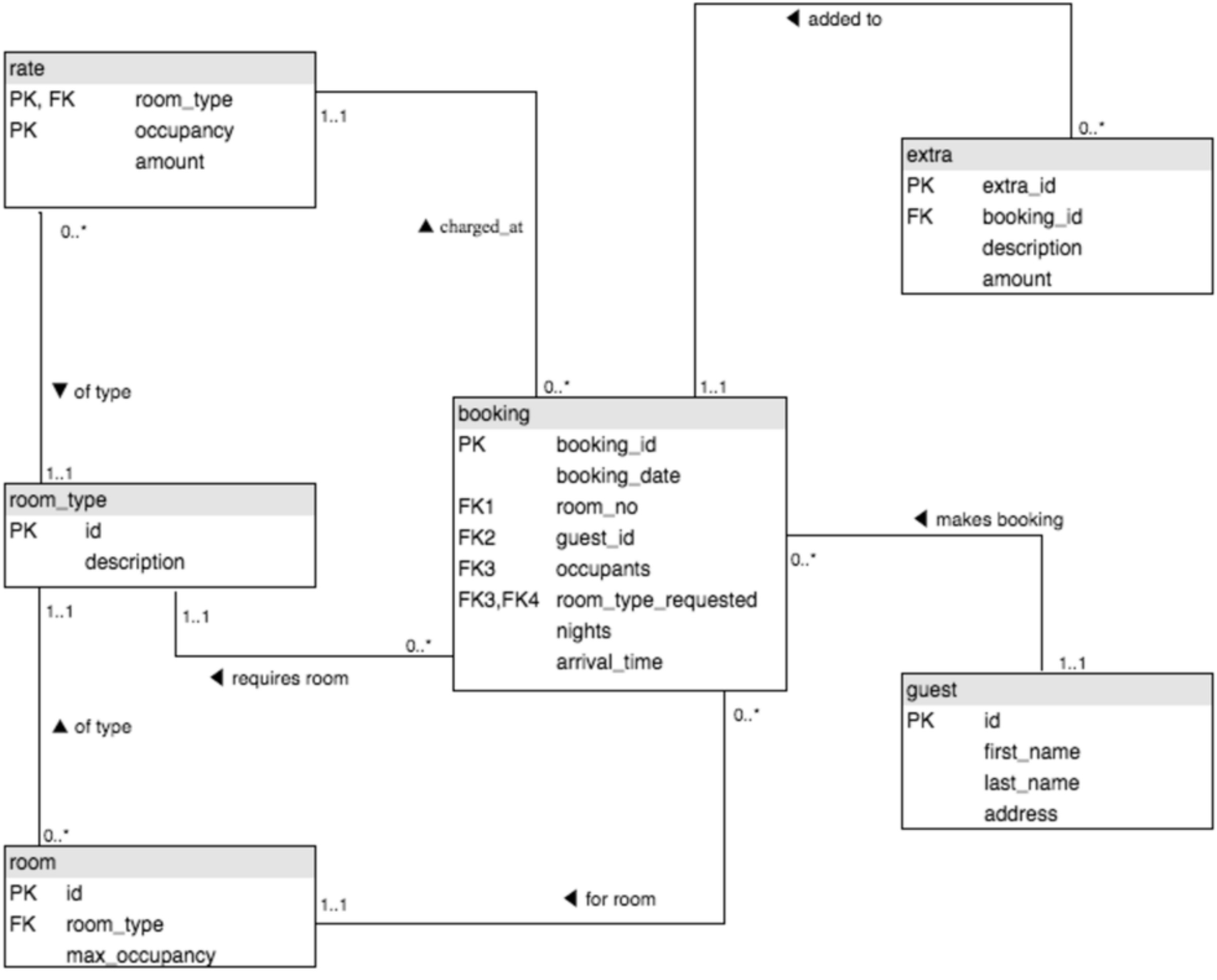
Tables can have many foreign keys

Used to link that table to many other tables

What is most likely the primary key?

Voter ID	First Name	Last Name	Turnout Score
1000001	Matt	Brems	0.96
1000002	Sam	Stack	0.43
1000003	Joseph	Nelson	N/A

Primary keys *must* be unique!


rate
PK, FK room_type
PK occupancy
amount
1..1
charged_at
0..*
added to
0..*
extra
PK extra_id
FK booking_id
description
amount
0..*
of type
1..1
1..1
booking
PK booking_id
booking_date
FK1 room_no
FK2 guest_id
FK3 occupants
FK3,FK4 room_type_requested
nights
arrival_time
1..1
makes booking
0..*
room_type
PK id
description
1..1
1..1
0..*
requires room
1..1
guest
PK id
first_name
last_name
address
0..*
of type
0..*
room
PK id
FK room_type
max_occupancy
1..1
for room

Each table has a schema

Voter ID	First Name	Last Name	Turnout Score
1000001	Matt	Brems	0.96
1000002	Sam	Stack	0.43
1000003	Joseph	Nelson	N/A

Column 1: “Voter Id” *(int)*
Column 2: “First Name” *(string)*
Column 3: “Last Name” *(string)*
Column 4: “Turnout Score” *(real)*

Your turn: Design an Uber database

Your turn: Design an Uber database

User ID
User Name
Driver ID
Driver Name
Ride ID
Ride Time
Pickup Longitude
Pickup Latitude
Pickup Location
Drop-Off Location
Drop-Off Latitude
Drop-Off Location Entity
Miles
Travel Time
Fare
CC Number

List the tables you would create

What fields would each table contain?

Remember, tables should be able to connect