GENERAL ASSEMBLY

DIFFERENT DATABASES

Matt Brems, Data Science Immersive

LEARNING OBJECTIVES

- Recognize common databases and know industry applications.
- Identify potential structure of SQL database in use-case.
- Describe what SQL and noSQL mean.

OPENING

- What is the maximum number of rows an Excel file can open at once?

- Databases are the standard solution for data storage and are much more robust than text, .csv, or .json files. Most analyses involve pulling data to and from a resource.

- This resource is, most commonly, a database.

INTRO TO RELATIONAL DATABASES

- Databases are computer systems that manage the storage and querying of data.
- Databases provide a way to organize data along with efficient methods to retrieve specific information.
- Databases also allow users to create rules that ensure proper data management and verification.
- Retrieval is typically performed using a query language – the most common is SQL. (Structured Query Language)

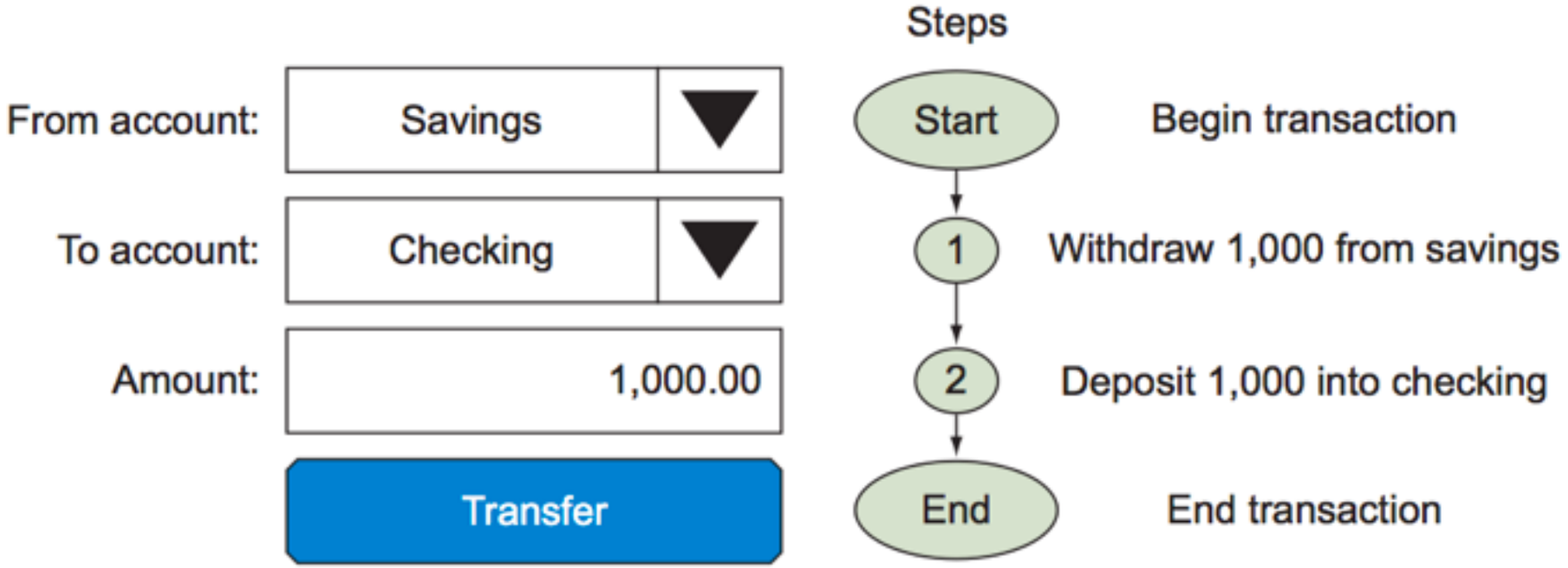
INDUSTRY EXAMPLE 1

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

- If this table was stored in a central server, what problems could arise?

TRANSACTIONAL INTEGRITY

- One unit of work performed against a database is called a “transaction.”
- This term generally represents any change in the database.



- This system must be resilient to any problems.
- ACID (Atomicity, Consistency, Isolation, Durability)

INTRO TO RELATIONAL DATABASES

- A relational database is a set of data organized in tabular (table-like) form with links between data entities or concepts.

- You can think of each table as similar to a single .csv file or a Pandas dataframe, with rows and columns.

- Each table typically has a primary key, which is a unique value per row serving as the identifier for that row.

- Each table can have many foreign keys, which link that table to other tables.

INDUSTRY EXAMPLE 1

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

- What is the likeliest primary key here?

INDUSTRY EXAMPLE 1

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

Voter ID	First Name	Last Name	2016 primary?	2016 general?
1000001	Matt	Brems	True	True
1000002	Sam	Stack	False	True
1000003	Joseph	Nelson	False	True

SCHEMA

- Each table has a “schema,” which is a set of rules for what goes in each table.

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)

UBER EXAMPLE

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

- List tables you would create.
- What fields would each contain?
- Remember that they must link to other tables.

ALTERNATIVE DATABASES

- Key-Value Stores: Very large and fast; similar to Python dictionaries but can be larger than your computer memory by relying on smart caching algorithms. Typically used for image stores, key-based file systems, object cache, systems designed to scale. (Popular ones include: Cassandra, Redis, memcachedb)

- NoSQL or Document Databases: Do not rely on a traditional table setup; often have nested data setups. Typically used for high-variability data, document search, Web content publishing. (Popular ones include mongodb and couchdb.)

- Time Series Databases / Graph Databases: Different structure.