Foundations of Data Science for Biologists

Introduction to SQL

BIO 724D

2024-NOV-12

Instructors: Greg Wray and Paul Magwene

Introduction to SQL

What is SQL?

SQL (**S**tructured **Q**uery **L**anguage) is a language for interacting with tabular data

- Specifically, it implements **relational** data structures

SQL is designed:

- for **powerful data queries** using a simple and compact syntax
- to enforce **data integrity** during data entry and updating
- for **highly efficient** search, sort, grouping, summarizing, etc. operations
- to be **massively scalable** (billions of rows, thousands of columns)

SQL is a **domain-specific language** (DSL)

- Designed to meet a specific set of needs
- Best-in-class for its intended purpose, bad-to-awful for most other purposes

SQL is designed to work with relational data structures

observations

seq	genus	species	date	location
1022	Colias	striatus	2007-06-14	Fairview Hotel
1023	Pycnonotus	tricolor	2007-06-14	Fairview Hotel
1024	Milvus	migrans	2007-06-14	Fairview Hotel
1032	Chalcomitra	amethystina	2007-06-14	Sheldrick Centre
1033	Pycnonotus	tricolor	2007-06-14	Sheldrick Centre
1050	Lamprotornis	superbus	2007-06-15	Lake Naivasha
1051	Lamprotornis	purpuroptera	2007-06-15	Lake Naivasha
1052	Scopus	umbretta	2007-06-15	Lake Naivasha
1053	Buteo	augur	2007-06-15	Hell's Gate NP
1054	Cisticola	marginata	2007-06-15	Hell's Gate NP

species

seq	genus	species	ssp	IUCN	familiar
1418	Pynonotus	leucogenys	1	LC	Himalayan Bulbul
1419	Pynonotus	barbatus	5	LC	Common Bulbul
1420	Pynonotus	tricolor	3	LC	Dark-capped Bulbul
1421	Pynonotus	capensis	1	LC	Cape Bulbul

locations

location	prov	country	clim	elev	geolocation
Everard Reserve	Vic	Australia	Csb	90	-37.68,145.49
Everglades NP	FL	USA	Aw	2	25.39,-80.63
Fairview Hotel	Nb	Kenya	Cfb	1715	-1.29,36.80
Faskrudsfjordur	Au	Iceland	ET	20	64.93,-14.01

Relational design removes redundant information by spreading it across tables:

(1) reduces errors, (2) simplifies updates, (3) saves space, (4) speeds up queries

Key features of relational data structures

Every table contains a **primary key** (PK) consisting of 1 column (or, occasionally, more)

- Every row must contain a value (no blanks or NULL values)
- Every value in this column must be unique (no duplicate entries allowed)

Every table should contain at least one **relation**

- Relation = a column that references a column in another table (typically its PK)
- Both columns must have the same data type; typically they contain overlapping values

The order of rows in a table is not consistent or stable

- This allows for highly efficient query and sort operations
- Retrieving data in a guaranteed sort order requires an explicit definition every time

SQL is limited to a small set of operations

Only four kinds of operations are permitted with relational databases

Create, **R**ead, **U**pdate, **D**elete (CRUD)

`CREATE` to create a new database, table, or relation; `INSERT` to add rows to a table

`SELECT` to query (retrieve information from) a database

`UPDATE` to change information in a table

`DROP` to remove a database, table, or relation; `DELETE` to remove rows from a table

SQL has some additional keywords, but most work with the above set

Clauses within statements: `LIMIT`, `WHERE`, `JOIN`, `GROUP BY`, etc.

Operators within statements: `=`, `>`, `AND`, `NOT`, `LIKE`, `BETWEEN`, etc.

Functions within statements: `MIN`, `MEAN`, `COUNT`, `DISTINCT`, etc.

`dplyr` and `Pandas` implement some SQL-like functions (and were inspired by it)

Practice database

Practice database: birding observations

Full implementation consists of 9 tables:

- 1 records field observations: the heart of the database
- 4 define taxonomy: species, genera, families, orders
- 3 define place: location, country, bioregion
- 1 defines context: trips/residences

Simplified version:

- Omits 2 tables, omits some columns from every table
- Includes >13.5K observations

Lophornis stictolophus
Spangled Coquette

Photo credit: Steve Gettle

Points to keep in mind about the database

Taxonomy follows the International Ornithological Congress (IOC)

- Scientific name: single word at each rank, follows rules of zoological nomenclature
- Familiar (common) name: single name, attempt to reconcile multiple synonyms
- Sequence of taxa within each rank: provides consistent ordering in lists

The **content** of the taxonomic tables differs by rank

- `orders` and `families`: complete list of taxa recognized by the IOC
- `genera` and `species`: only taxa that have been observed
- `subspecies`: not in a separate table, but recorded in `observations` table

Primary keys are based on how the database is queried and updated

- `orders`, `families`, and `genera`: taxon name; `species`: IOC sequence (binomen is hard!)
- `locations` and `trips`: brief name; `observations`: temporal sequence