Midterm Review & SQLite

The Midterm…

- You all passed!
- I returned points lost due to errors on my part.
  - If you wrote “ETL” instead of “Extract, Transform, Load” on 4/4: +2
  - If you wrote “causal” on 11/2: +2
- Too easy questions: 5, 7, 10, 13-20, 25, 29
- Some of the questions were unintentionally tricky: 8, 21, 22, 24, 26.

Question 1 – Correct: 7

Question 1

Match each of the 4 V's of Big Data wtih the closest meaning.

- ✔ __2__ The quality of data.
- ✔ __1__ The speed at which data are generated.
- ✔ __3__ Multiple sources of data.
- ✔ __4__ The size of the dataset.

1. Velocity
2. Veracity
3. Variety
4. Volume

Question 2 – Correct: 7

Question 2

Which of the following are steps in a typical data science workflow?

- ✓ ☐ Machine learning.
- ✓ ☑ Data modeling.
- ✓ ☑ Collect data.
- ✓ ☑ Communicate the results.
- ✓ ☐ Cloud computing.
- ✓ ☑ Define the problem.
- ✓ ☑ Data exploration.

Question 3 – Correct: 7

Question 3

Match each data type with its closest description.

- ✔ __1__ Data that assume a data model.
- ✔ __4__ Data about data.
- ✔ __3__ Data that can be parsed to fit a data model.
- ✔ __2__ Data that do not fit a pre-defined model.

1. Structured Data

2. Unstructured Data

3. Semi Structured Data

4. Metadata

Question 6 – Correct: 5

Question 6

In python, a dict comprehension can be used for the following:

- [x] Modify dictionary values.
- [x] Convert a dictionary to a list. <- list(dict.items())
- [x] Filter a dictionary.
- [x] Order a dictionary. <- Dictionaries are unordered.

Question 8 – Correct: 0

Question 8 **0 / 2 points**

Your https GET request returned a 5XX status code. Your request is:

- ✖ Received. <- We know it was received because it got a response and did not time out.
- ✓ In process. <- Not in process because the request failed. This is 1XX.
- ✓ Valid. <- This is in the 5XX description.
- ✓ Requires further action. <- This is 3XX.
- ✓ Contains bad syntax. <- This is 4XX.

Question 8 – Correct: 0

Question 9 – Correct: 7

Question 9 **2 / 2 points**

Hypothesis generation results from inferential data science problems.

- ○ True
- ✓ ● False <- Hypothesis generation results from exploratory data science problems.

Question 11 – Correct: 0

Question 11 **4 / 4 points**

Determining whether your data fit a larger population is the goal of this type of data science problem. ___inferential___ ✔(50 %) Determining the input variable that changes an output variable is the goal of this type of data science problem. ___causal___ ✖ **(causasl)** <- Sorry.

Question 12 – Correct: 4

Question 12 **2 / 2 points**

Beautiful Soup is used to make HTML GET requests in python.

- ○ True
- ✓ ● False

<- requests.get

Question 21 – Correct: 2

Question 21 **2 / 2 points**

Age in years is an example of which kind of data?

- Ordinal.
- Interval.
- ✓ Ratio. <- Continuous, preserves order, interval of known size, has true zero.
- Nominal.

Question 22 – Correct: 3

Question 22 **2 / 2 points**

Classifier algorithms always generate this type of data?

- ✓ Nominal. <- Classifiers generate a label.
- Ratio.
- Interval.
- Ordinal.

Question 23 – Correct: 7

Question 23 **2 / 2 points**

Sentinel values and dummy variables are similar in that they are both used to represent another value.

- ✓ (●) True
- ( ) False

Question 24 – Correct: 1

Question 24 **4 / 4 points**

Using pandas, sentinel values can be specified in which parameters?

- [x] usecols <- Specifies columns to use.
- [x] true_values <- Specifies sentinel values representing True
- [x] na_filter <- Omits null records
- [x] delim_whitespace <- Specifies whitespace as a separator.
- [x] false_values <- Specifies sentinel values representing False

Question 26 – Correct: 2

Question 26 **6 / 6 points**

Match the python expression with a valid description.

✔ __1__ np.nan	**1**. Null
✔ __3__ set([ ])	**2**. String
✔ __1__ None	**3**. False

Question 27 – Correct: 7

Question 27 **2 / 2 points**

The pandas apply() function can be used to clean string elements in a data frame column using a function.

- ✔ (●) True
- ( ) False

Question 28 – Correct: 6

Question 28 **2 / 2 points**

The python map() function can not be used to clean string elements in a list using a function.

- ( ) True
- ✓ (•) False

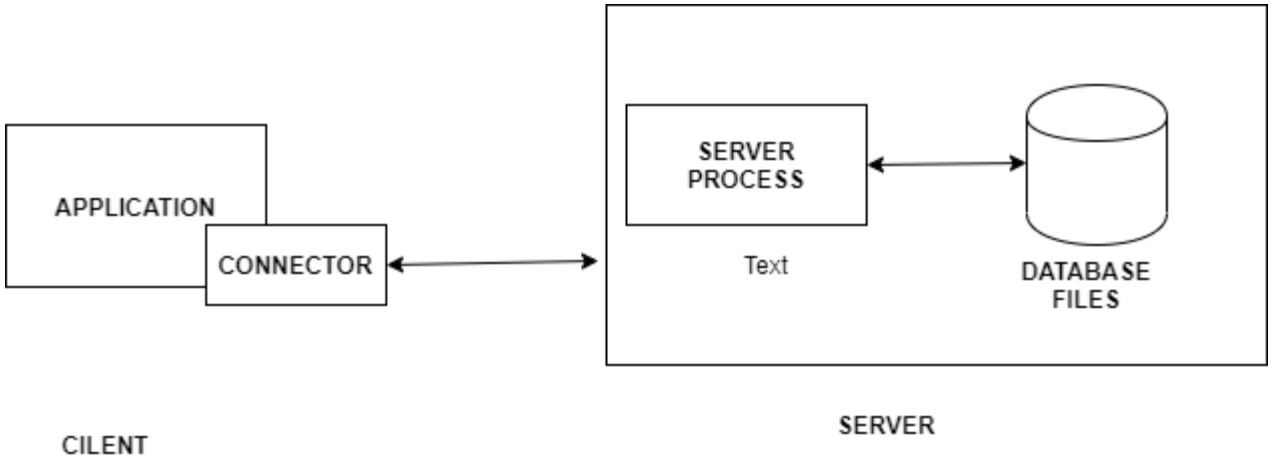
SQLite

- A “light” version of MySQL:
  - Serverless
  - Self-contained
  - Zero-configuration
  - Transactional
  - Single-Database

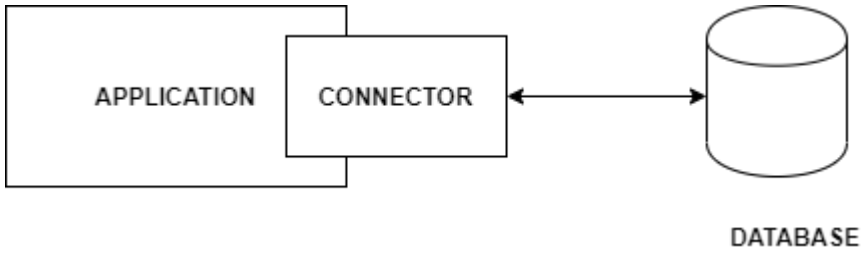
Ref: https://www.geeksforgeeks.org/python-sqlite/

Typical vs. SQLite

- Typical RDBMS Configuration:



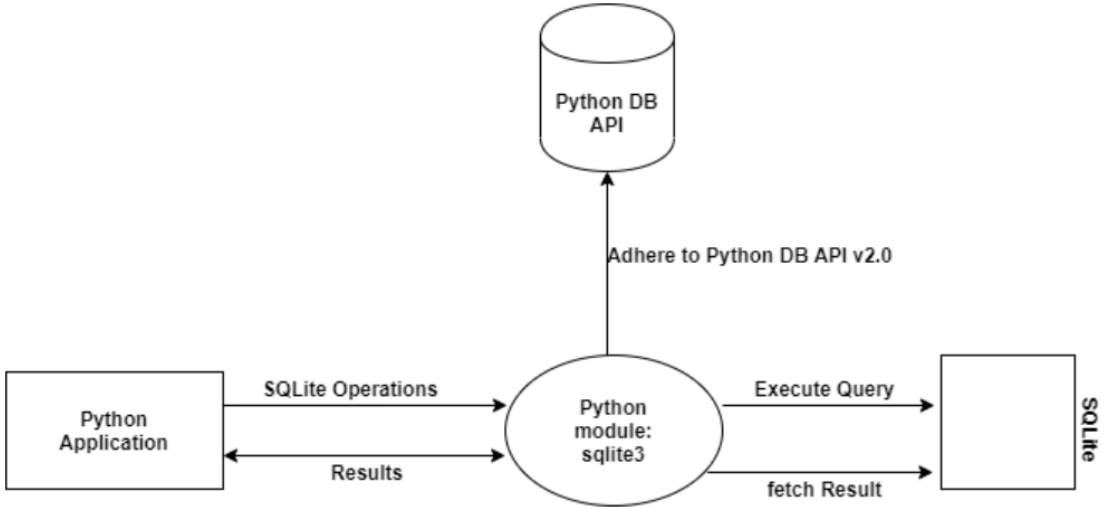
- SQLite Configuration:



- No server. Applications read/write/interact with files on disk stored in the DB

SQLite

- Self-contained: all you need is SQLite.
- Zero-configuration: no setup or administration
- Transactional: ACID
- Single-Database: A single connection to access multiple files.



SQLite Storage Classes

Storage Class	Value Stored	Python Datatype
NULL	NULL	None
INTEGER	Signed integer (1,2,3,4,5, or 8 bytes)	Int
REAL	Floating point (8 byte)	Float
TEXT	Text string (UTF-8, UTF-16BE, or UTF-16LE)	Stir
BLOB	Binary format input	Bytes

SQLite commands

- SELECT
- WHERE
- LIMIT
- DELETE
- ORDER BY
- GROUP BY
- AND
- OR
- MIN
- MAX
- AVG
- SUM
- COUNT

Ref: https://www.dataquest.io/blog/sql-basics/

Homework 3

I’ll release it before next week. The goal is for you to think about how you should structure the data for your project.

It starts with breaking your data down into relationship tables.

Here’s a handy reference: https://opentextbc.ca/dbdesign01/chapter/chapter-8-entity-relationship-model/