LECTURE 8

SQL

Relational Databases and Querying a Table

Data 100/Data 200, Fall 2021 @ UC Berkeley
Fernando Pérez and Alvin Wan
(content by Alvin Wan, Anthony D. Joseph, Allen Shen, Josh Hug, John DeNero, Joseph Gonzalez)

LECTURE 8

Introduction

Motivation, Definition, and Takeaways

Data 100/Data 200, Fall 2021 @ UC Berkeley
Fernando Pérez and Alvin Wan
(content by Alvin Wan, Anthony D. Joseph, Allen Shen, Josh Hug, John DeNero, Joseph Gonzalez)

Drawbacks of a file

What is a “Database”?

Takeaways

How to Try SQL

Why your **file** will **fail** you.

Can’t scale

Unreliable

Unoptimized

Unstructured

Why your **database** will **help** you.

Scalable

Reliable

Optimized

Structured

Drawbacks of a CSV

What is a “Database”?

Takeaways

How to Try SQL

A **database** is an organized collection of data.

A **database management system** (DBMS) is a software system that **stores**, **manages**, and **facilitates access** to one or more databases .

SQL is a language for managing data in a DBMS.

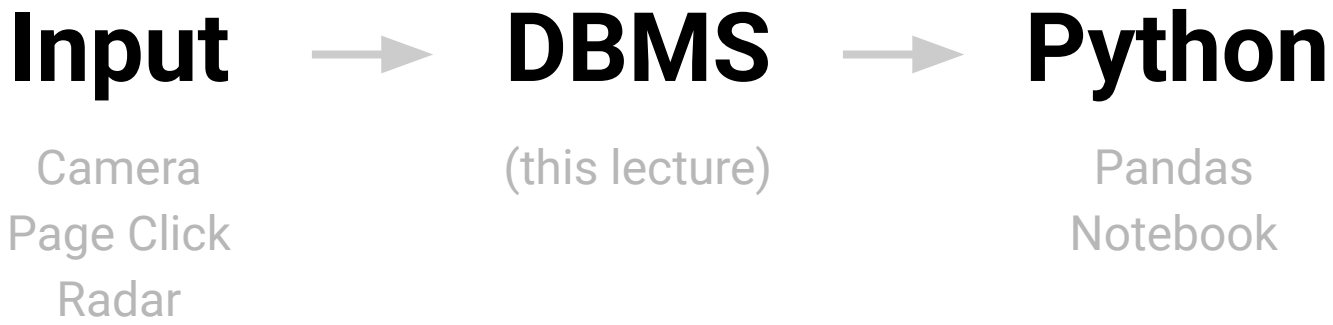
Drawbacks of a CSV

What is a “Database”?

Takeaways

How to Try SQL

Dataflow at Work



What you should focus on.

Not memorize *SQL syntax*
Understand what SQL can do
Write clean SQL Code
Debug SQL Queries

Drawbacks of a CSV
What is a “Database”?
Takeaways

How to Try SQL

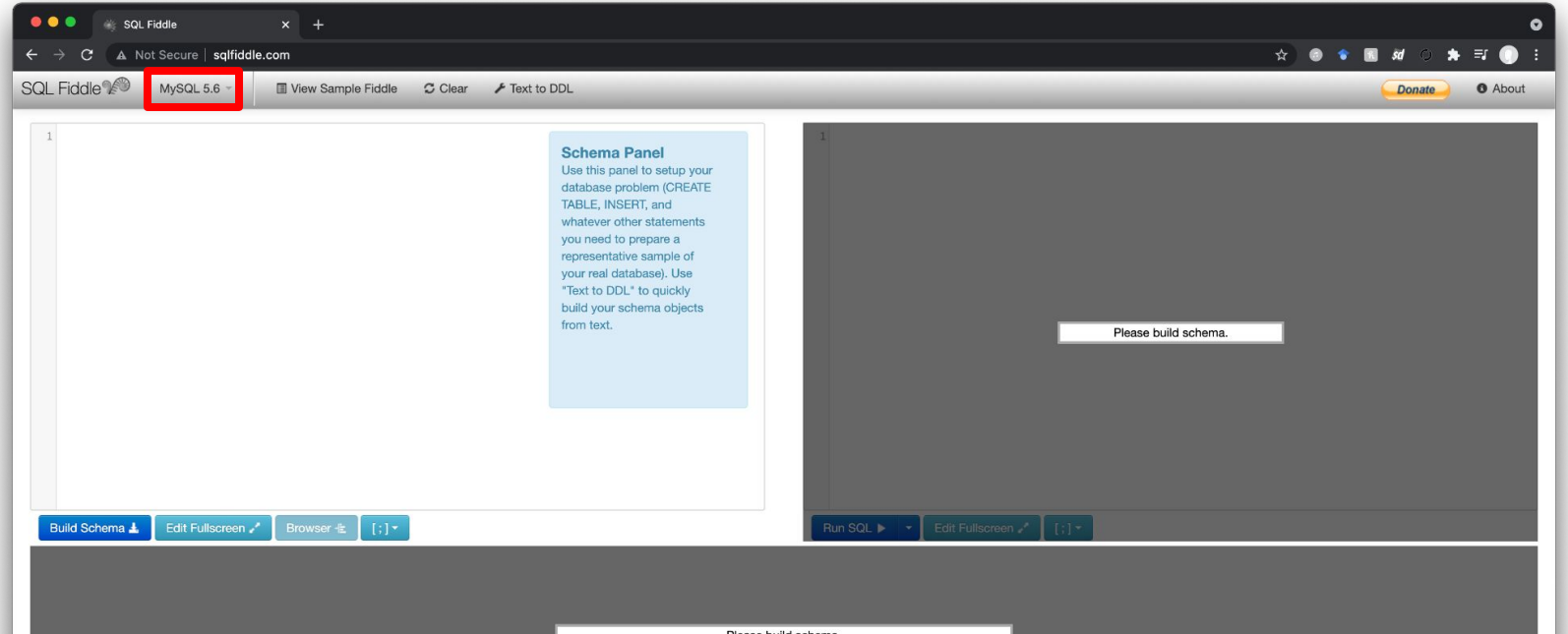
Most popular DBMSs

What we'll use.
Easy to setup.
Missing features.

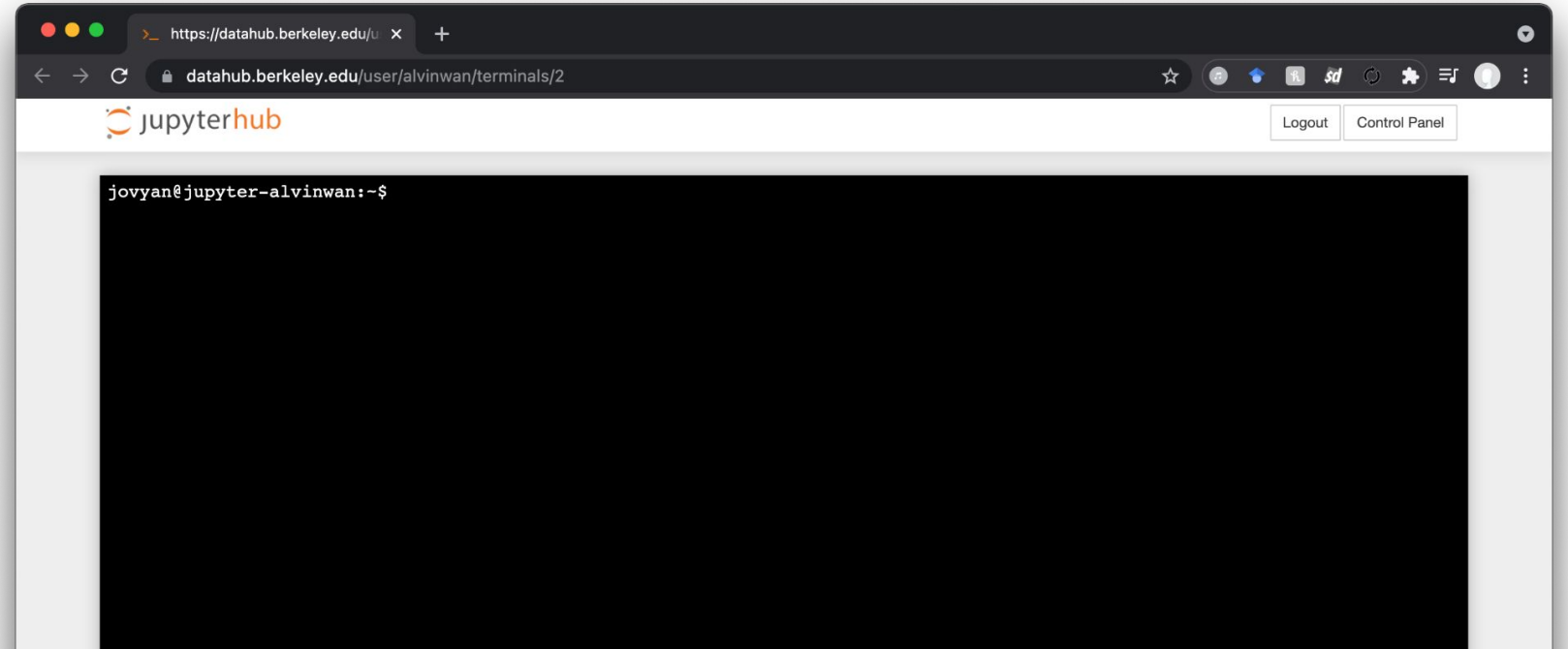
359 systems in ranking, August 2020

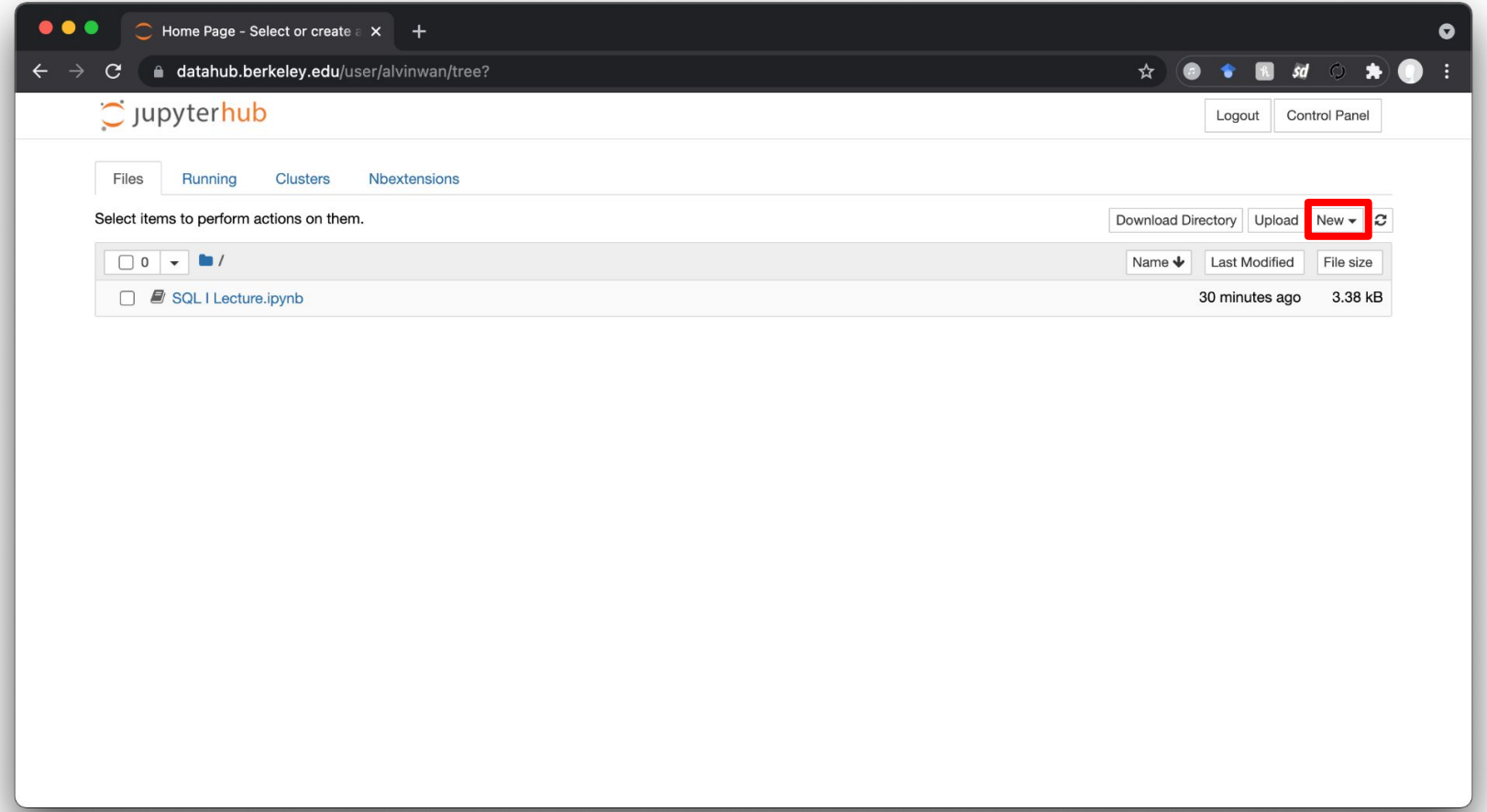
Rank Aug 2020	Rank Jul 2020	Rank Aug 2019	DBMS	Database Model	Score Aug 2020	Score Jul 2020	Score Aug 2019
1.	1.	1.	Oracle	Relational, Multi-model	1355.16	+14.90	+15.68
2.	2.	2.	MySQL	Relational, Multi-model	1261.57	-6.93	+7.89
3.	3.	3.	Microsoft SQL Server	Relational, Multi-model	1075.87	+16.15	-17.30
4.	4.	4.	PostgreSQL	Relational, Multi-model	536.77	+9.76	+55.43
5.	5.	5.	MongoDB	Document, Multi-model	443.56	+0.08	+38.99
6.	6.	6.	IBM Db2	Relational, Multi-model	162.45	-0.72	-10.50
7.	↑ 8.	↑ 8.	Redis	Key-value, Multi-model	152.87	+2.83	+8.79
8.	↓ 7.	↓ 7.	Elasticsearch	Search engine, Multi-model	152.32	+0.73	+3.23
9.	9.	↑ 11.	SQLite	Relational	126.82	-0.64	+4.10
10.	↑ 11.	↓ 9.	Microsoft Access	Relational	119.86	+3.32	-15.47
11.	↓ 10.	↓ 10.	Cassandra	Wide column	119.84	-1.25	-5.37

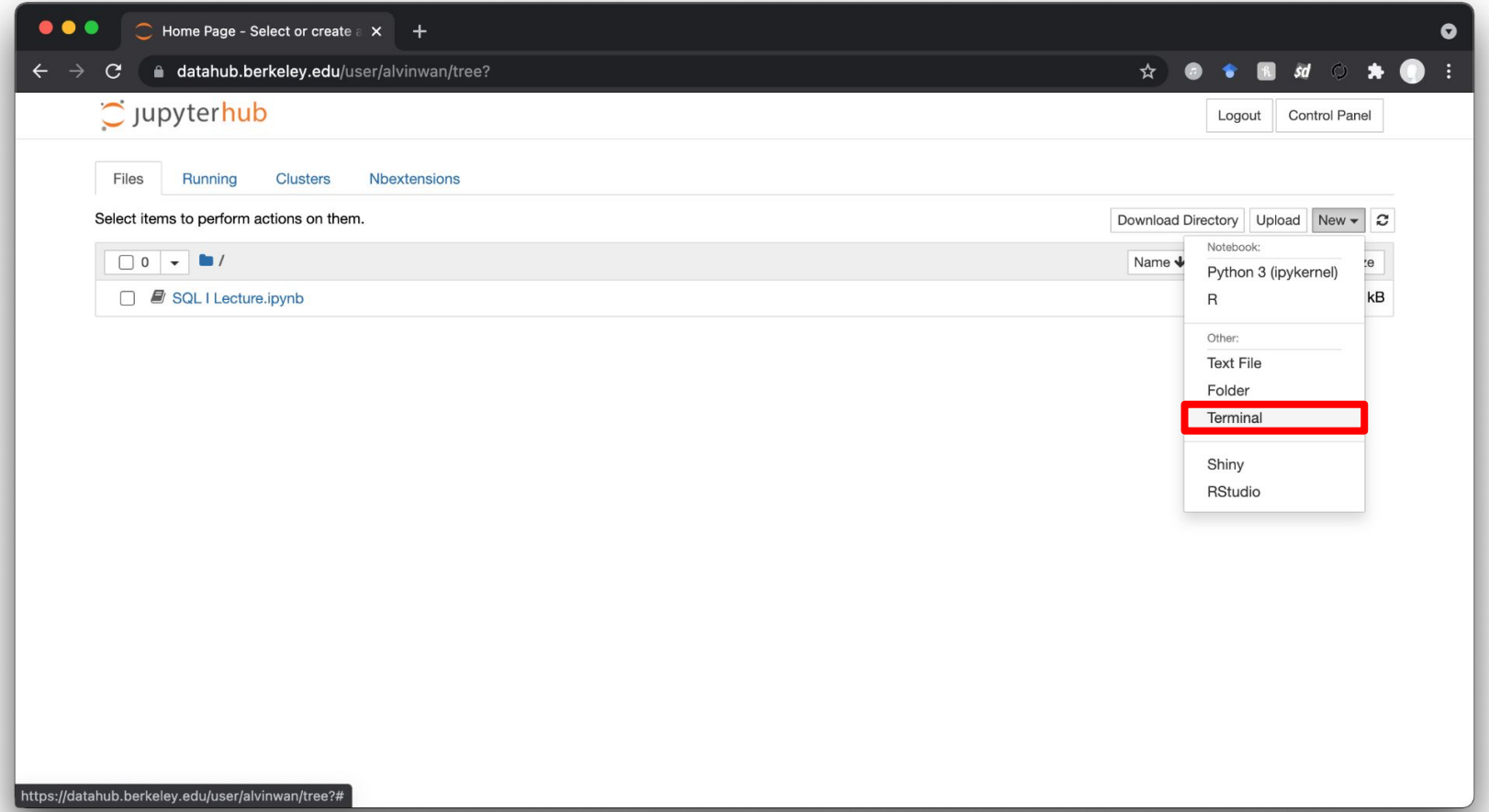
sqlfiddle.com

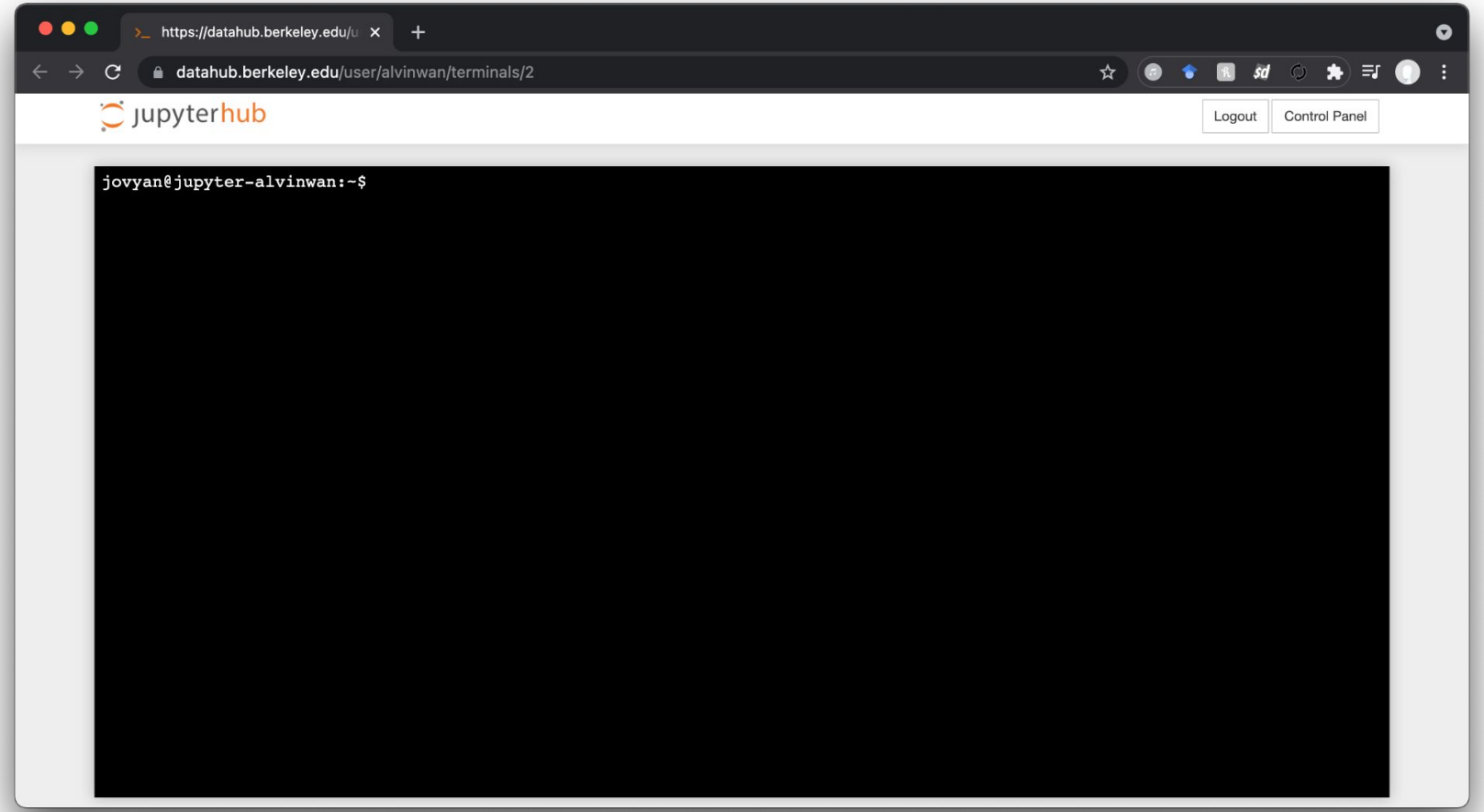


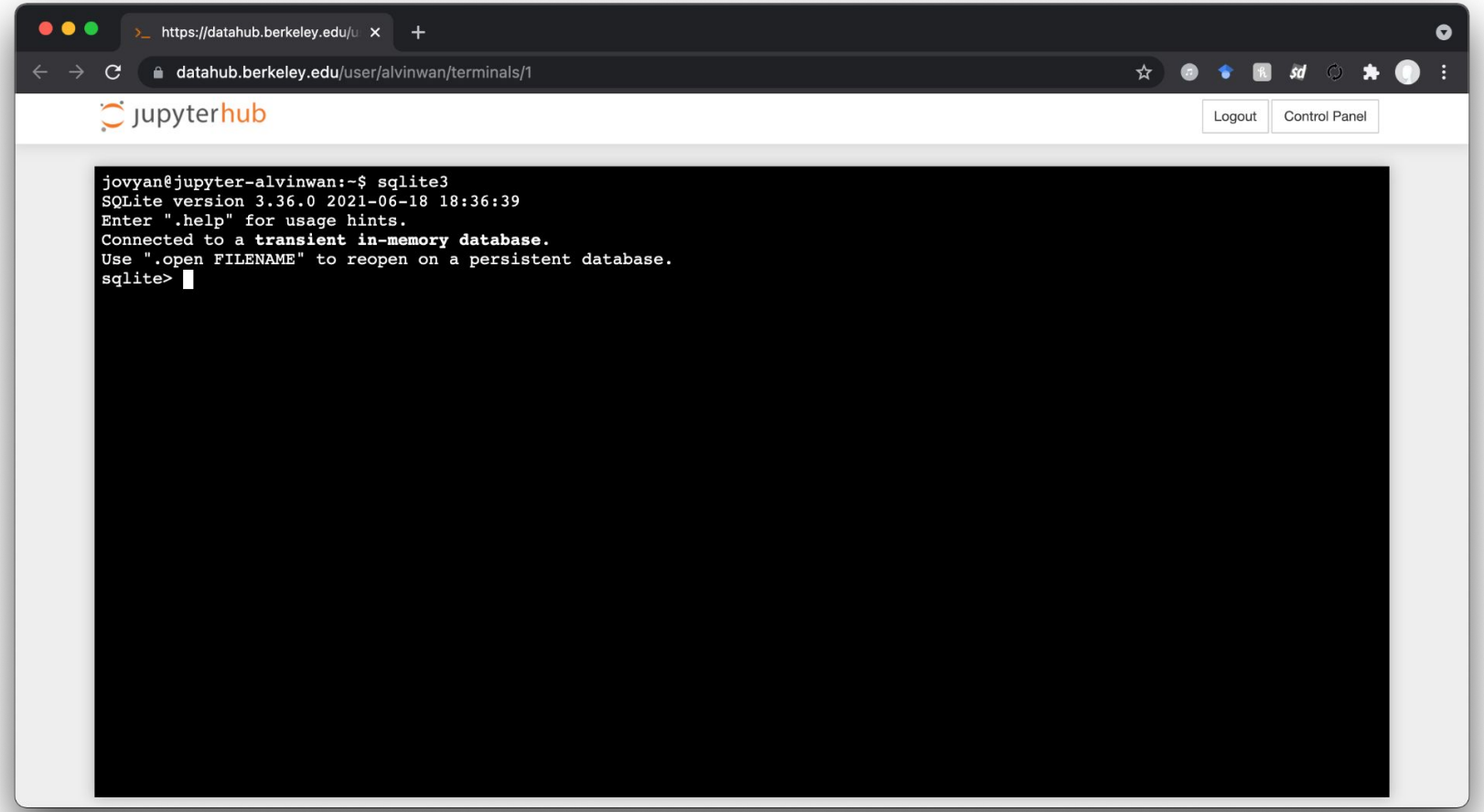
datahub.berkeley.edu terminal



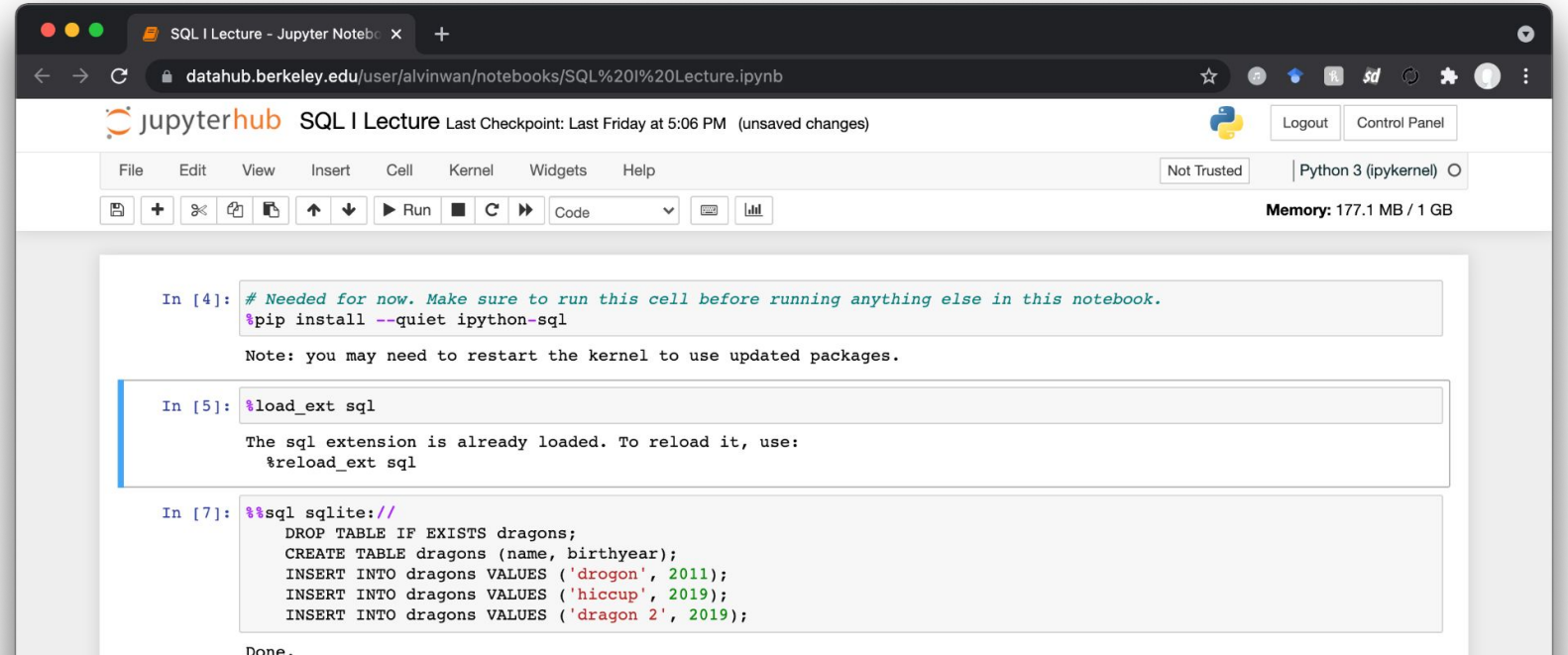

Home Page - Select or create
datahub.berkeley.edu/user/alvinwan/tree?
jupyterhub
Logout
Control Panel
Files
Running
Clusters
Nbextensions
Select items to perform actions on them.
Download Directory
Upload
New
0
/
Name
Last Modified
File size
SQL I Lecture.ipynb
30 minutes ago
3.38 kB

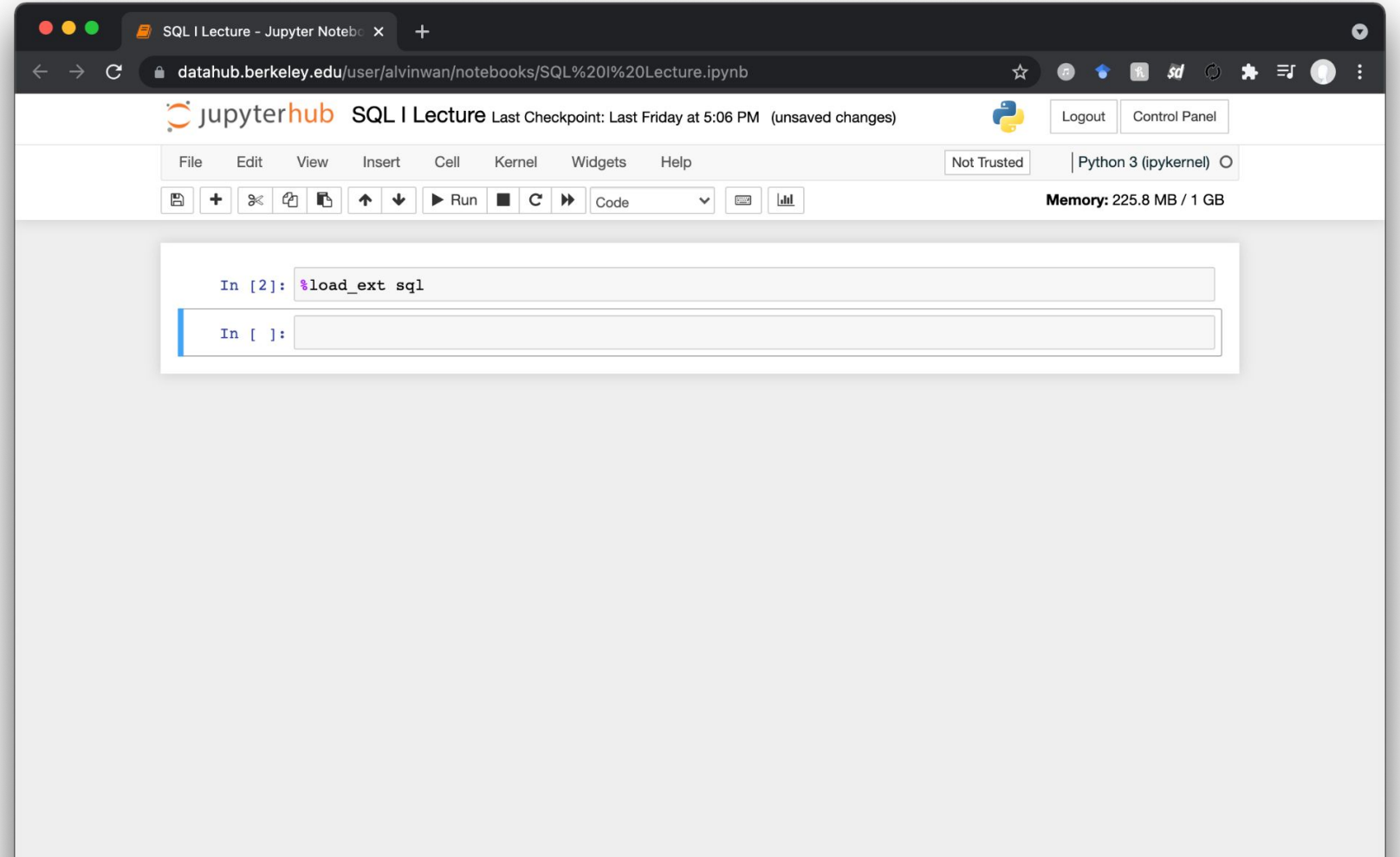

Home Page - Select or create a
datahub.berkeley.edu/user/alvinwan/tree?
jupyterhub
Logout
Control Panel
Files
Running
Clusters
Nbextensions
Select items to perform actions on them.
Download Directory
Upload
New
0
/
Name
SQL I Lecture.ipynb
kB
Notebook:
Python 3 (ipykernel)
R
Other:
Text File
Folder
Terminal
Shiny
RStudio
https://datahub.berkeley.edu/user/alvinwan/tree?#


https://datahub.berkeley.edu/u
datahub.berkeley.edu/user/alvinwan/terminals/2
jupyterhub
Logout
Control Panel
jovyan@jupyter-alvinwan:~$


https://datahub.berkeley.edu/u
datahub.berkeley.edu/user/alvinwan/terminals/1
jupyterhub
Logout
Control Panel
jovyan@jupyter-alvinwan:~$ sqlite3
SQLite version 3.36.0 2021-06-18 18:36:39
Enter ".help" for usage hints.
Connected to a transient in-memory database.
Use ".open FILENAME" to reopen on a persistent database.
sqlite>

datahub.berkeley.edu notebook




SQL I Lecture - Jupyter Notebo
datahub.berkeley.edu/user/alvinwan/notebooks/SQL%20I%20Lecture.ipynb
jupyterhub
SQL I Lecture
Last Checkpoint: Last Friday at 5:06 PM (unsaved changes)
Logout
Control Panel
File
Edit
View
Insert
Cell
Kernel
Widgets
Help
Not Trusted
Python 3 (ipykernel)
Run
Code
Memory: 225.8 MB / 1 GB
In [2]: %load_ext sql
In [ ]:

TAKEAWAY

Databases improve **reliability**, **scalability**, and **cleanliness** of your data. Leverage this with SQL.

LECTURE 8

Terminology

Terminology and Examples

Data 100/Data 200, Fall 2021 @ UC Berkeley
Fernando Pérez and Alvin Wan
(content by Alvin Wan, Anthony D. Joseph, Allen Shen, Josh Hug, John DeNero, Joseph Gonzalez)

Table

Examples

Column or **Attribute** or **Field**

Row or **Record** or **Tuple**

type TEXT, PK	legs INT, >=0	weight INT, >=0
Corgi	4	10
T-Rex	2	12000
Penguin	2	10

Schema:
Name,
Type,
Constraint

Animal

Table or **Relation**

PRACTICAL TIP

For simplicity and convenience, keep table names singular (and camel-case).

stackoverflow.com/a/5841297/4855984

Types	Constraints
INT	NOT NULL
REAL	DEFAULT
TEXT	UNIQUE
BLOB	PRIMARY KEY
NULL	CHECK
~~ENUM~~	

PRACTICAL TIP

By convention, queries use all caps for SQL keywords.

softwareengineering.stackexchange.com/q/261684

Table

Examples

Animal

type TEXT, PK	**legs** INT, >=0	**weight** INT, >=0
Corgi	4	10
T-Rex	2	12000
Penguin	2	10

- No duplicate animal **type**s
- No negative **legs**
- No negative **weight**

Member

athlete TEXT, PK	**esport** TEXT, PK	**skill** INT
Danny	Warzone	10
Jane	Starcraft	1000
Jane	Warzone	100000

No duplicate (**athlete**, **esport**) tuples *but* duplicate **athlete** (Jane) *or* duplicate **esport** (Warzone) allowed

PRACTICAL TIP

Make the primary key an integer, for efficiency and convention.

bit.ly/3hKmnbw

Clothing

id INT, PK, AUTOINC	**sku** TEXT, UNIQUE	**name** TEXT, NOT NULL
1	92183	blouse
2	23012	jeans
3	57603	polo

- **id** auto-populated, incremented
- No duplicate **sku**s
- No empty **name**s

PRACTICAL TIP

Add `created_at` and `updated_at` columns for maintainability. Make them self-update.

TAKEAWAY

SQL offers **types** and **constraints** to enforce data cleanliness.

LECTURE 8

Conclusion

Takeaways

Data 100/Data 200, Fall 2021 @ UC Berkeley
Fernando Pérez and Alvin Wan
(content by Alvin Wan, Anthony D. Joseph, Allen Shen, Josh Hug, John DeNero, Joseph Gonzalez)

Summary

- Databases are
  - (more) reliable
  - scalable
  - optimized
  - structured - Know all the SQL constraints you can impose.
- SQL is powerful. Know all the questions you can answer with SQL.
- SQL to preprocess. Python to postprocess.
- Keep in mind your SQL tips for clean, optimized queries.