Big Data

Objectives

- Introduction to the processes of creating and working within databases which deal with large data
- To help solidify the foundations of MySQL and SQL syntax
- To learn how to create MySQL schemas and import large amounts of external data into a database

Pop Quiz!

See 11-popQuiz in class repo

What is ‘big data’?

Extremely large data sets that may be analyzed computationally to reveal patterns, trends, and associations, especially relating to human behavior and interactions

One of the advantages of MySQL over other server systems is that it can easily handle large datasets

What is a CSV file?

comma separated values

Activity: Examining the Dataset

See TopSongs.csv

- Columns
  - artist name
  - song name
  - year
  - raw popularity score for the entire world
  - raw popularity scores for
    - US
    - UK
    - Europe
    - Rest of world

* "raw score" numbers reflect the "total value of music industry sales", where a higher raw score indicates a greater volume of sales

Activity: Examining the Dataset

With a partner, discuss how you might go about creating a database for this dataset

id	artist	title	year	rawps	rawUS	rawUK	rawEU	rawET

Does your server save your SQL queries?

null

What is saved?

Ch-ch-ch-changes...

What if we wanted/needed to setup an identical database on another server?

Demo: Setting up Schemas and Planting Seeds

- schema.sql
  - Used to store database creation code
- seeds.sql
  - Used to store statements for inserting data into tables

CTRL + C
CTRL + V

Activity: Preparing the Database (20 min)

See Slack for instructions...

Somebody Slack me a schema.sql

Review: Preparing the Database

Demo: Importing Data

- Nothing up my sleeve: `SELECT * FROM Top5000;`
- Import TopSongs.csv
- Not quite magic
- Presto!

Activity: Importing & Working with Big Data (50 min)

See Slack for file and instructions…

RTFM https://www.w3schools.com/sql/sql_groupby.asp

&& https://www.w3schools.com/sql/sql_between.asp

Review: Importing & Working with Big Data

```
var query = "SELECT artist FROM top5000 GROUP BY artist HAVING count(*) > 1";
```

GROUP BY groups elements with shared values together and then allows us to use the HAVING count(*) >1 statement to determine if there are multiples within that group

```
var query = "SELECT position,song,artist,year FROM top5000 WHERE position BETWEEN ? AND ?";
```

BETWEEN ? AND ? allows us to select information between a specific range

RTFM

https://dev.mysql.com/doc/refman/5.7/en/mysql-indexes.html

&& https://atech.blog/viaduct/mysql-indexes-primer

When dealing with big databases, it is likely that you will have to work with two or more datasets that are related, but which have some degree of separation between them

A table stores information in rows and columns.

A database is a collection of related tables.

Activity: Two Tables Are Better Than One (60 min)

See 14-TwoTables for files and instructions…

Review: Two Tables

Homework