WEEK 2 LAUNCH GROUP PROJECT

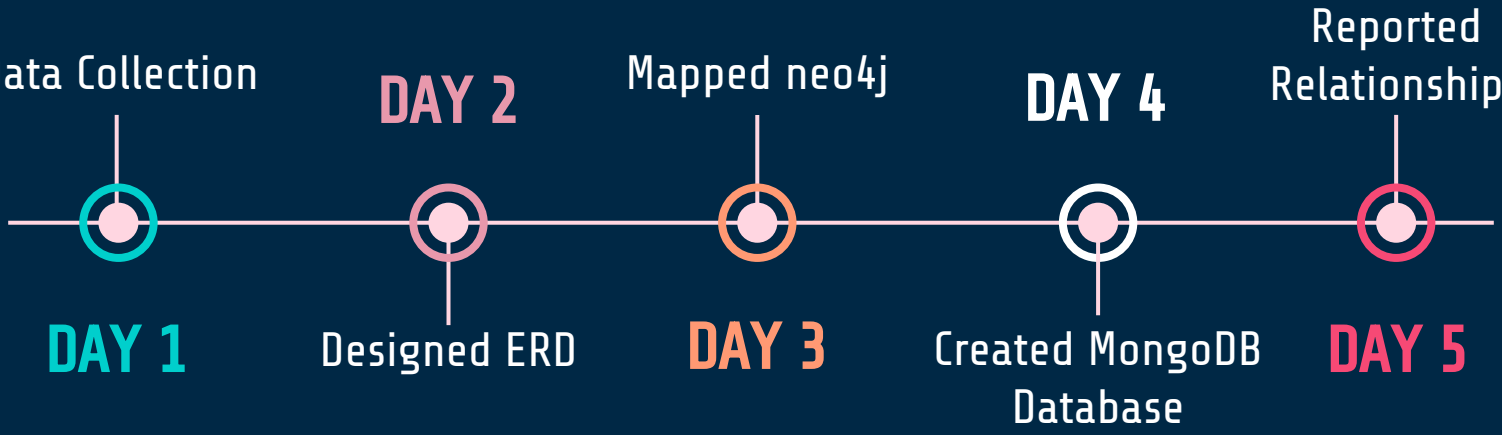
Drops of Jupyter

Adi Pillai, Donald Cooper,
Julia Duarte, Steven Burke,
Sydney Ploeger

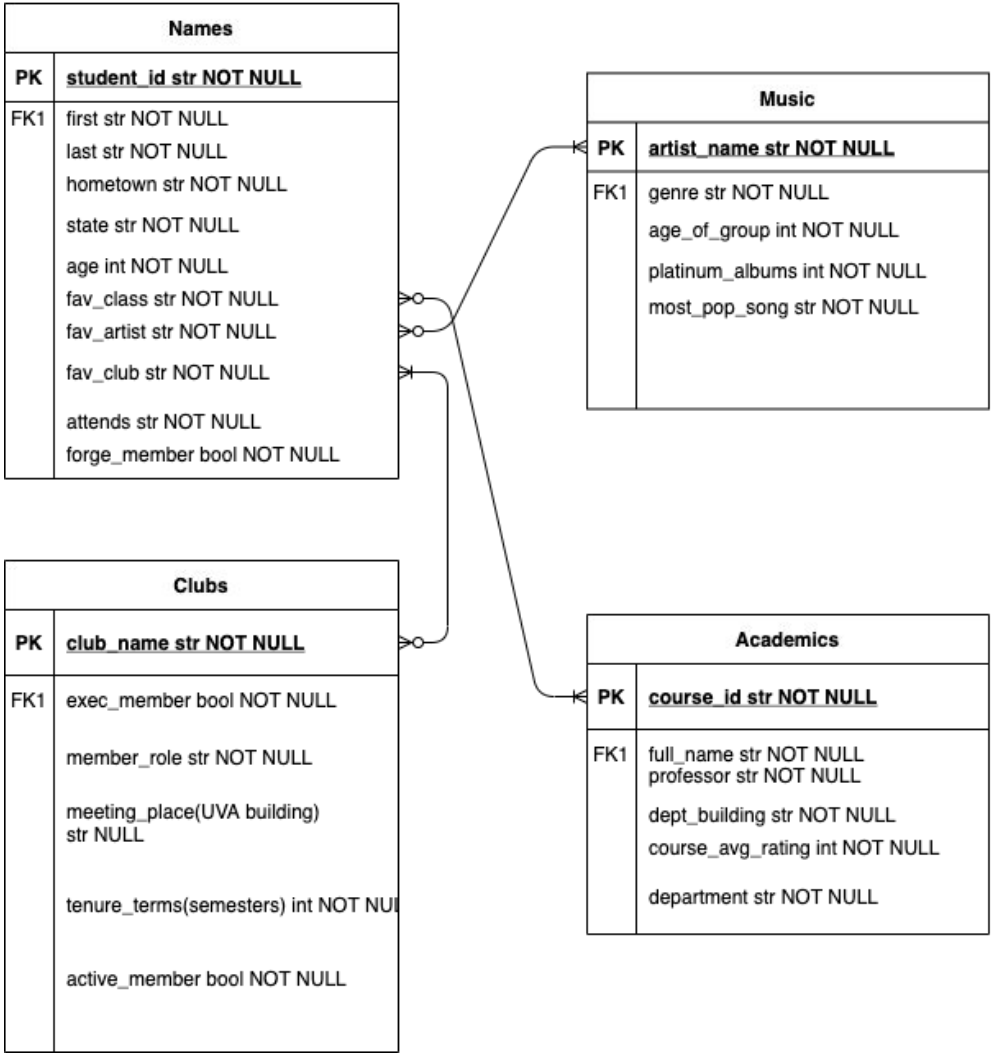
Goals for Presentation

- Process
- Preliminary Visualizations
- Database Comparisons
  - Neo4j
  - Mongo
  - MySql
- Final Evaluation

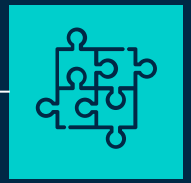
OUR PROCESS


Data Collection
DAY 1
DAY 2
Designed ERD
Mapped neo4j
DAY 3
DAY 4
Created MongoDB Database
Reported Relationships
DAY 5

PRELIMINARY ERD



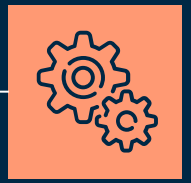
DATABASE EXPLORATION



1

NEO4J

Graphing visualizations of key-value relationships.



2

MONGODB

Nesting based on shared relationships between tables.

3

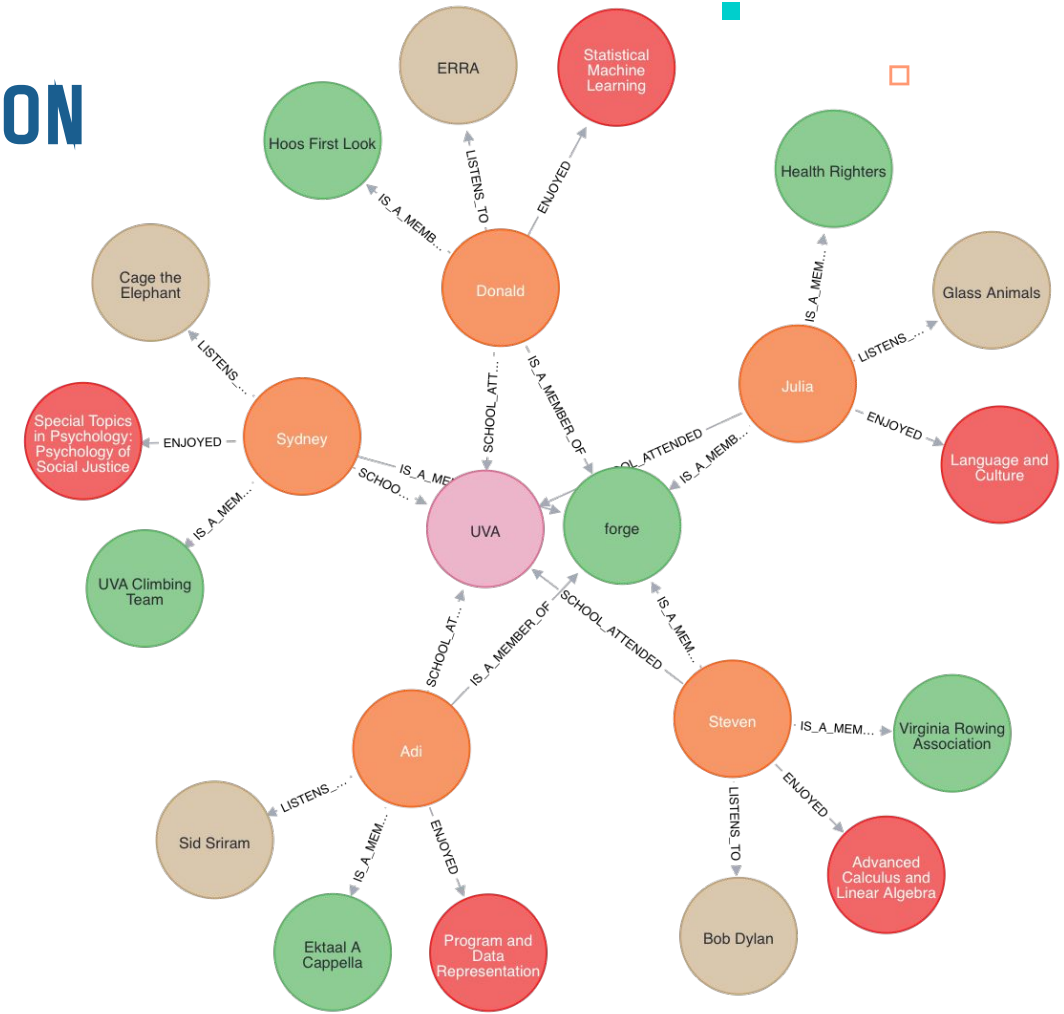
MYSQL

Querying unique relationships via RDBMS

NEO4J

1

VISUALIZATION



Advantages

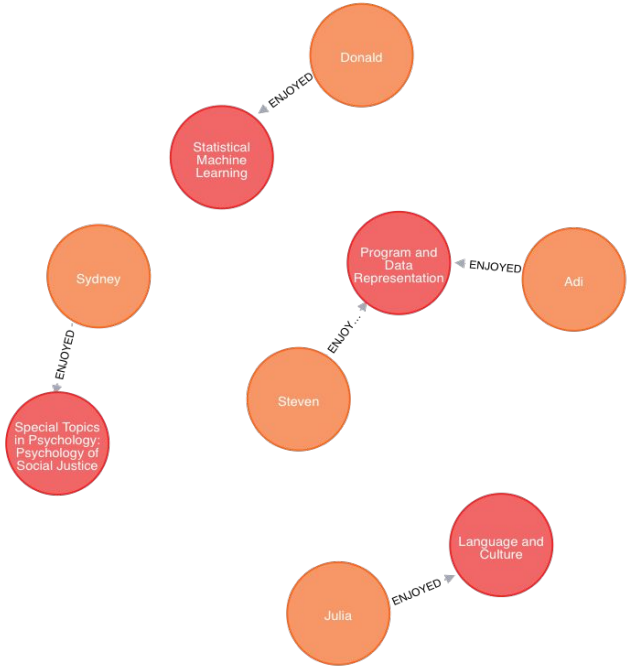
- Easier visualization between connections.
- Performance during/after new data integration.
- Flexibility during industry changes.
- Agility prevents clunky development pipelines.

Disadvantages

- Limited EDA and analytics hampers BI practices
- Not optimized for processing large datasets (“Bigdata”)
- Some relationships are difficult to map depending on database architecture
- More difficult aggregations

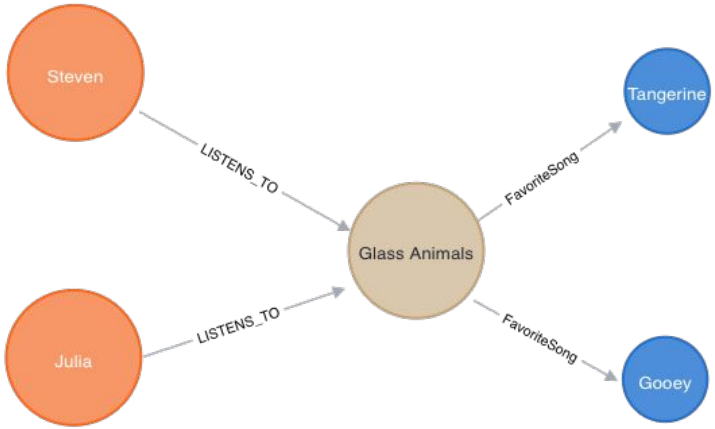
Example Query 1

```
neo4j$ MATCH (a)-[:ENJOYED]→(b) return a,b
```



- Quickly visualizes all data points
- Can make comparisons for categorical data
- No bijective relationships

Example Query 2

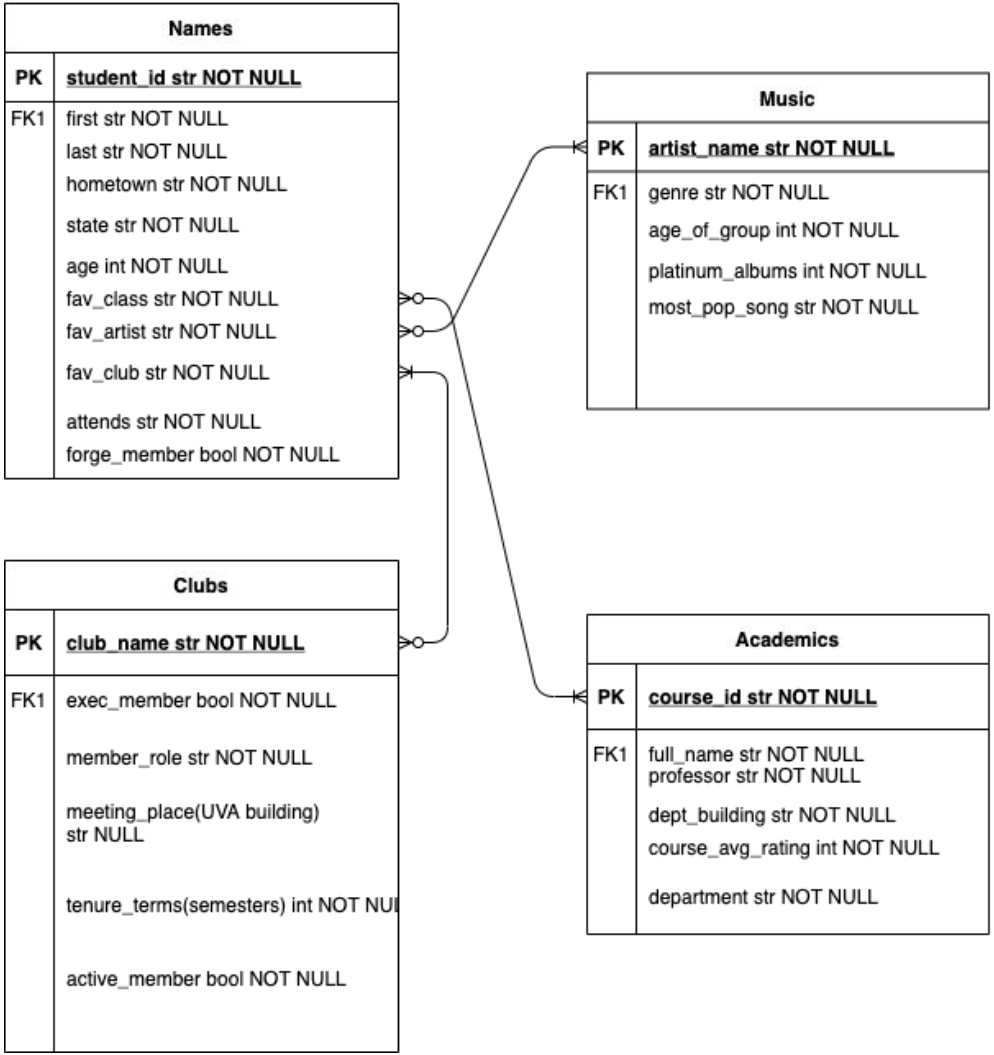


- Ownership of properties can be difficult
- Sometimes hard to map related but nuanced data

```
neo4j$ MATCH (a)-[:LISTENS_TO]→(b)-[:FavoriteSong]→(c) RETURN a,b,c
```

MONGODB

2



PRELIMINARY ERD MONGODB

VISUALIZATION

- CS 2150 Document from the Courses Collection
- Will be nested in our primary collection

VISUALIZATION – Primary Collection

This document contains the previous document from the Courses Collection

Example document from primary collection contains nested documents from sub-collections

Advantages

- Easy to track all possible data for a given document
- Easily understood subdocuments
- Instituting new data is facilitated well by the document structure

Disadvantages

- Sometimes difficult to compare data
- No standardization among the documents in a collection
- Learning curve not very intuitive

Example Query

Which students are younger than 20 and have a favorite artist older than 20?

Able to filter on the fields of the document as well as the nested fields, which is unique to MongoDB

Example Aggregation

```
/**
 * _id: The id of the group.
 * fieldN: The first field name.
 */
{
  _id: "$attends",
  avg_rating: {"$avg" : "$fav_class.course_avg_ratin
}
```

```
_id: "UVA"
avg_rating: 4.266
```

What is the average course rating for our favorite classes?

Aggregations not as robust as in SQL or Node because the documents are independent from each other

Can still perform aggregations with documents in a collection if they have the same fields

MYSQL

3

VISUALIZATION

#	column_name	data_type	character_set	collation	is_nullable	column_default	extra	foreign_key
1	student_id	varchar(255)	utf8mb4	utf8mb4_0900_ai_ci	NO	NULL	EMPTY	EMPTY
2	first	text	utf8mb4	utf8mb4_0900_ai_ci	YES	NULL	EMPTY	EMPTY
3	last	text	utf8mb4	utf8mb4_0900_ai_ci	YES	NULL	EMPTY	EMPTY
4	hometown	text	utf8mb4	utf8mb4_0900_ai_ci	YES	NULL	EMPTY	EMPTY
5	state	text	utf8mb4	utf8mb4_0900_ai_ci	YES	NULL	EMPTY	EMPTY
6	age	bigint	NULL	NULL	YES	NULL	EMPTY	EMPTY
7	fav_artist	varchar(255)	utf8mb4	utf8mb4_0900_ai_ci	YES	NULL	EMPTY	music(artist_name)
8	fav_class	varchar(255)	utf8mb4	utf8mb4_0900_ai_ci	YES	NULL	EMPTY	academics(course_id)
9	fav_club	bigint	NULL	NULL	YES	NULL	EMPTY	clubs(index)
10	attends	text	utf8mb4	utf8mb4_0900_ai_ci	YES	NULL	EMPTY	EMPTY
11	forge_member	tinyint(1)	NULL	NULL	YES	NULL	EMPTY	EMPTY

index_name	index_algorithm	is_unique	column_name
fav_artist	BTREE	FALSE	fav_artist
fav_class	BTREE	FALSE	fav_class
fav_club	BTREE	FALSE	fav_club
PRIMARY	BTREE	TRUE	student_id

The foreign key represents relationships between the "name" table's columns and another table's columns (i.e. "music", "academics", and "clubs").

fav_artist	fav_class	fav_club
Sid Sriram	CS 2150	4
ERRA	STAT 4630	3
Glass Animals	ANTH 2400	1
Bob Dylan	MATH 2315	5
Cage the Elephant	PSYC 4500	2

Advantages

- Easily customizable queries
- More accessible comparisons
- Most familiar type of Database (relational databases)
- Can have a greater variety of data

Disadvantages

- Queries and joins can be computationally demanding.
- Not as easy to visualize relationships within data.
- Stricter datatype requirements.

Example Query 1

```
pd.read_sql('''SELECT * FROM names
               JOIN music ON names.fav_artist = music.artist_name
               JOIN academics ON names.fav_class = academics.course_id
               JOIN clubs ON names.fav_club = clubs.index
                   WHERE age <= 20 AND hometown = "Virginia Beach"''', con = cnx)
```

student_id	first	last	hometown	state	age	fav_artist	fav_class	fav_club	attends	...	dept_building
jod8esu	Julia	Duarte	Virginia Beach	VA	18	Glass Animals	ANTH 2400	1	UVA	...	Brooks Hall
sep2nb	Sydney	Ploeger	Virginia Beach	VA	20	Cage the Elephant	PSYC 4500	2	UVA	...	Gilmer Hall

Which students are located in VA and are 20 years old or less?

Querying categorical information is simple with smaller data.

Example Query 2

```
pd.read_sql('''SELECT genre, platinum_albums,
               SUM(platinum_albums)
               FROM music
               GROUP BY genre
               ORDER BY SUM(platinum_albums) DESC;''', con = cnx)
```

	genre	platinum_albums	SUM(platinum_albums)
0	Folk	12	17.0
1	Rock	9	17.0
2	Pop	1	14.0
3	Progressive House	9	9.0
4	Rap	5	5.0
5	Electronic	3	3.0
6	Hip-Hop	3	3.0
7	Folk Rock	2	2.0

Which genre has the most platinum albums?

We can see the immediate advantages of query aggregations. However, we lose the visual appeal via table printouts.

Best Choice of Database

Perfect for our rather small dataset

Well suited for comparing categorical variables

Neo4j

Best at visualizing relationships (students connections)

Most visually appealing

Other Considerations for Mongo and MySQL...

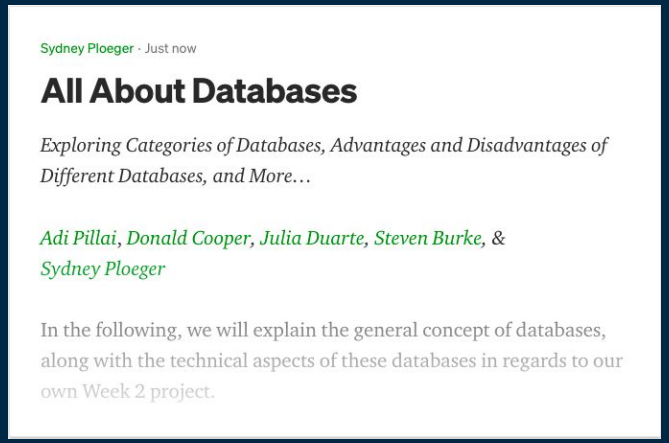
MongoDB

- Nesting more information within existing columns.

MySQL

- Calling more BI/Analytics queries for data summaries and/or aggregation.

Additional Things!


Sydney Ploeger · Just now

All About Databases

Exploring Categories of Databases, Advantages and Disadvantages of Different Databases, and More…

Adi Pillai, Donald Cooper, Julia Duarte, Steven Burke, & Sydney Ploeger

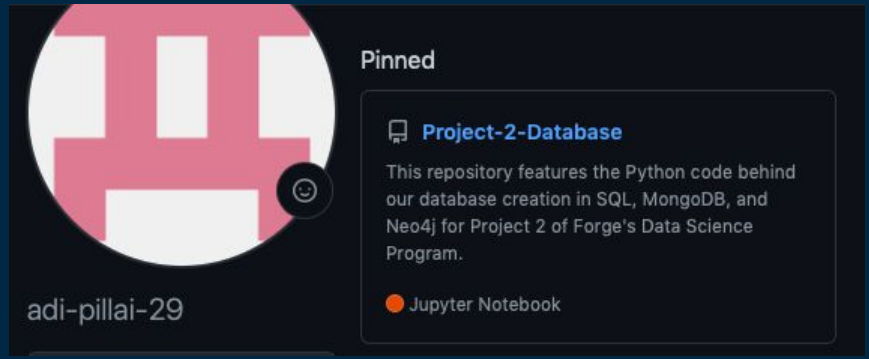
In the following, we will explain the general concept of databases, along with the technical aspects of these databases in regards to our own Week 2 project.

Medium Article

https://sep2nb.medium.com/all-about-databases-ca47cb3b784e

GitHub Repository

https://github.com/adi-pillai-29/Project-2-Database.git



Thank you!

Questions?