CHATDB FOR LARGE-SCALE ENTERTAINMENT DATASETS

Team Details

Team members background and skills

- Quynh Tran: Background in Business,
    - My name is Quynh Tran and my undergraduate is Business Administration. I have experience in managing and making business decisions based on statistics. I just transferred to Data Science and am doing my master's in person at USC.
- Kevin Bui:
    - My name is Kevin Bui, and I am currently a DEN Student majoring in Computer Science Data Science for my Master of Science. I work full-time as an Application Support Specialist at Center for Creative Leadership. This is a fully remote position, and I work heavily within SQL Server to maintain live data in many different applications and API's. This involves performing data fixes, creating/executing stored procedures, and creating reports for data analysis. I studied at Azusa Pacific University with a Bachelor of Science in Computer Science. I am comfortable with Python, Java, SQL, CSS, PHP, React, and Javascript.
- Riya Berry:
    - My name is Riya Berry, and I am currently an in-person Masters in Applied Data Science student at USC. I completed my BA in Data Science at University of California, Berkeley, and then worked for 2 years as a Data Analyst at a market research consulting firm called Protagonist. I have significant experience in building and developing ML systems, NLP, statistical modeling, and data mining in Python/R/SQL. I am also currently conducting Computer Science Research for the Norman Lear Center at USC Annenberg. As a researcher, I am developing a web-interface platform that allows users to search for key terms within a large-scale database of film and television scripts. Through this role, I have gained increased exposure to front and back-end development, AWS EC2 and ElasticBeanstalk, and distributed database management.

Project requirements

In this project, our task is to create a NLI (natural language interface) that interfaces with at least 3 distinct SQL and NoSQL database systems. We will design an NLI that performs CRUD operations using natural language inputs. In other words, users can write out requests to the NLI in natural/plain English – such as “how many rows does this dataset contain?” and other questions to explore the kinds of data that exist in the database. Users can also query our database to find out what tables exist, apply functions such as *match* and *sort*, and modify data in the database – i.e., “delete all data from the year 2022” or “update John’s location to California.”

Our code will leverage Large Language Model (LLM) APIs to translate user’s natural language requests into the appropriate SQL/NoSQL query. Depending on the user’s query, our NLI should also be able to create, update, and delete data. As a three-person group, we will also be using full-stack development to create a web platform for users to interact with ChatDB. This platform will allow users to ask plain language questions about the data, process requests in the back-end, and translate the results into output that is easily understandable to non-technical users. In other words, this platform will serve as an intelligent “ChatGPT” website users can visit for modifying/exploring/and querying our database (ChatDB).

For this project, we must make a proposal, submit progress reports, and demo our project to the rest of the class. The majority of our work lies in the implementation, which must be completed before our in-class demo.

Planned Implementation

1. Development Tools & Environment

- **Version Control:** GitHub (for code sharing)
- **Database Types:**
    - **SQL:** MySQL
    - **NoSQL:** options
        - MongoDB
        - Amazon DynamoDB (integrates with EC2, S3)
        - Cassandra (Apache)
        - *Consideration:* Firebase (with REST API integration)

2. Large Language Model (LLM) API Integration

- **Primary Choice:** Mistral API (via HuggingFace)
- **Decision Needed:** Determine if other LLM APIs might be a better fit through exploration

3. Dataset Selection & Preparation

- **Industry:** Entertainment/Media
- **Potential Large-Scale Datasets for Database:**

    - Kaggle (for large, easy-to-access datasets)
    - Movie Reviews
    - Concerts (Scraped from LiveNation, TicketMaster)
    - Google Dataset Search (for CSVs, Excel Sheets)
- **Tasks:**
    - Scraping and cleaning the dataset

4. NLI Development

- **Goal:** Develop a web app for the NLI interface
- **Task:**
    1. Figure out which type of LLM API we're using
    2. Front end for inputting responses and outputting results (need to have a beta NLI interface)
    3. Exploration of database
        - First get NLI working on SQL database, then the NoSQL database

Team Responsibilities:

- Rachel: project & timeline management
    - Primary role in developing and finalizing our written reports, as well as creating presentation for live demo
    - SQL server management
- Riya:
    - Researching/setting up LLM API to manage natural language inputs
    - Distributed data management
    - SQL server management
- Kevin:
    - NoSQL server management
    - Front/back end development

Timeline

Task #	Task Name	Assigned Team Member(s)- Team Responsibility	Start Date	End Date	Status	Notes - Planned Implementation
1	Project Kickoff & Planning	Everyone	02/06	02/12	C…	☑ ~~Initial brainstorming, project scope finalized~~ ☑ ~~Research plan on implementing ChatDB~~ ☑ ~~Identify database source~~ ☐ https://www.kaggle.com/datasets/shiivvvaam/most-watched-movies-and-tv-shows ☑ ~~Set up Git~~ ☑ ~~Set up all the tools we need and do research how we are going to connect and use them~~ ☑ ~~LLM~~
2	Submitting Proposal	Ri…	F…	Date	C…	100 POINTS
2	Dataset Preparation	Riya		2/12	C…	☑ ~~Data cleaning using Pandas, export to CSV & JSON~~ ~~Use pandas for cleaning, export to CSV~~ ☑ ~~Find a way to make data structured for SQL and NoSQL during cleaning~~ ☑ ~~Convert to JSON for NoSQL~~
3	SQL Database Implementation	Riya, Rachel		2/17	C…	☑ ☑ ~~MySQL server set up (Riya and Rachel)~~
4	NoSQL Database Implementation	Kevin		2/17	C…	☑ ~~Set up MongoDB/DynamoDB/Cassandra~~ ☑ ~~Mysql and postgresql~~
5	Natural Language Processing (LLM API	Riya		2/24	C…	☑ ~~Integrate Mistral API via Hugging Face to NLI~~

	Integration)					
6	Backend Development	Kevin		3/5	C…	☑ ~~Develop API endpoints for SQL/NoSQL queries~~ ☑ ~~Set up MySQL server, test queries~~ ☑ ~~Integrate Hugging face- open ai 3.5 (rachel)~~ ☑ ~~Firebase setup~~

7	Mid-term Progress report	All	M…	Mar …	C…	☑ ~~Write the report - rachel is writing- verify some information with Riya and Kevin~~
						☑ ~~Tested out the modification request with Langchain -> Worked~~
						
						Task needed completed 03/07/2025
						☑ ~~Complete these parts in the report~~
						  ☑ ~~What has been implemented so far? What are the next steps?~~
						  ☑ ~~Describe your approach to converting natural language into database queries.~~
						☑ ~~Ask kevin for his part of NoSQLI to confirm~~
						☑ ~~Copy the timeline to the mid-progress report~~
						☐ Riya Berry
						☑ ~~Duplicates in the database for tv shows (each season)~~
						  ☑ ~~Choose entry with the highest watch time~~
						☑ ~~Join in data about tv show episode and season count (TMDB)~~
						  ☑ ~~Have to join at least two tables~~
						  ☑ ~~Check that "titles" match when joining these tables~~
						  ☐

8	Tools Setup Check+ Checking NLI			3/12	In…	**Task Needed completed before March 12th 2025**
						☐ Set up ChatDB with Langchain
						  ☐ Team Members need to make sure How to run it on local machine:
						  ☐ Langchain, OpenAI,

	(Necessity of having ec2)					streamlite and Langchain to run it ☐ All run in one file ☐ Kevin Bui Could you upload the code to Google Collab for test run for Riya and Rachel in our loca machine ☐ Build UI for query submission & visualization ☐ Kevin - turn data to sql ☐ Kevin idea: all run the demos, by uploading the code to run - run it locally ☐ Integrate our specific database ☐ Everyone: ☐ Push everything to github (cleaned, nosql, setting up) ☐ Riya Berry ☐ Clean duplicate tv shows (titles should be unique) ☐ Create a JSON file where the key is title ☐ Clean new dataset about tv show episode / season count
9	Testing and Debugging			03/19	N…	☐ Unit testing for query handling
10	Testing and Debugging			3/26	N…	☐ Report structure, challenges & solutions
11	Testing and Debugging	P…	D.	Apr …	N…	
12	Testing and Debugging	P…	D.	Apr …	N…	
14	Testing and Debugging	P…	D.	Apr …	N…	☐ Prepare for the Demo

14	DEMO DAY	P…	D.	Apr …	N…	☐ Demo (inclass, 4/21 and 4/23, 10 points): ☐ Give a live demo of your project. All project members should be present during the demo, presenting his/her contribution. If you're absent, we will assume that you have not contributed to the implementation of the project
	FINAL TOUCH_UP	P…	D.	Apr …	N…	☐ We should touch up on Final Report one more time before the MAY 09
15	FINAL REPORT	P…	M…	Date	N…	☐ the final report should be comprehensive, details your design and implementation, and your learning experiences. ☐ Implementation (due on your demo time, 60 points): note your project should be fully implemented before the demo. You should include in your final report a link to Google drive where you will upload your project codebase and documentations. Make sure you give access to your project folder.
		P…	D.	Date	N…	