DS5110 Final Project Iteration 02

Jinyu Li, Yichen Zhang, Richard Chen

October 2025

1 Project Kickoff

1.1 Specific Goals and Expected Outcomes

Goals:

- Build a chatbot system that automatically retrieves and organizes user reviews and comments from a data set of approximately 1,200 records.
- Use Python for backend automation to handle data ingestion, cleaning, and processing.
- Use SQL for reliable data storage, query handling, and retrieval of chatbot responses.
- Develop a TypeScript/JavaScript frontend where users can interact with the chatbot and view relevant review information.
- Deploy the entire system on AWS, ensuring scalability, reliability, and smooth integration across all components.

Expected Outcomes:

- A functional chatbot capable of pulling and displaying review data directly from the dataset upon user request.
- A clean, interactive web interface for users to browse or query reviews seamlessly.
- An automated backend pipeline that processes the dataset and updates SQL records with minimal manual intervention.

1.2 Project Scope Definition

- Backend data automation using Python.
- Review storage and query handling with SQL.
- Frontend chatbot interface with TypeScript/JavaScript.
- AWS setup for deployment and hosting.
- Integration between all three layers (backend, database, and frontend).

1.3 Key Deliverables for Each Project Phase

Phase 1 Data Review and Preparation: Review dataset structure, identify missing values, clean data, and prepare for SQL mapping.

Phase 2 Database Design (ERD Creation): Design ERD, define keys, and create SQL schema.

Phase 3 Backend Automation (Python): Build scripts to automate data ingestion, connect to SQL, and verify data flow.

Phase 4 SQL Implementation: Develop and test queries, validate performance, and confirm correct data retrieval.

Phase 5 Frontend Integration (TypeScript/JavaScript): Build chatbot interface, connect to backend, and deploy on AWS.

1.4 Major Milestones and Deadlines

Phase 1 Data review and preparation (7–10 days) → Clean dataset approved.

Phase 2 ERD creation (7–10 days) → Database schema finalized.

Phase 3 Backend automation (7–10 days) → Python automation verified.

Phase 4 SQL implementation (7–10 days) → Validated SQL database ready for frontend.

Phase 5 Frontend and AWS deployment (7–10 days) → Chatbot deployed and fully functional.

1.5 Team Capabilities and Alignment

- Most team members have experience in one or more phases.
- Work is divided based on each member's strengths (Python, SQL, or TypeScript/JavaScript).
- Each member contributes effectively to ensure balanced progress across all phases.

1.6 Dataset Availability

- A dataset of approximately 1,200 records is available for this project.
- It will be used for all testing and development.
- No external datasets are required for this iteration.

2 Team Discussions

2.1 Core Skills

- **Jinyu Li:** Strong data pre-processing skills using Python pandas. Skilled in SQL database and SQL query design. Experiences in Power BI/Tableau for visualization and presentation.
- **Yichen Zhang:** Strong Front-End development skills, including HTML, CSS, JavaScript, and React for the web interface. Familiar with RESTful API design and the techniques to connect front-end and back-end.
- **Richard Wang:** Skilled in Python and Flask/FastAPI for backend development. Experienced in data cleaning, data pipelines, and different analytical models. Familiar with deploying small-scale web applications to AWS.

2.2 Members' expertise

- **Data Preprocessing -** Jinyu & Richard: will clean raw data, remove noise, and generate analysis ready dataset.
- **SQL Database Creation -** Jinyu: will design schema, set up relations, and load processed dataset.
- **Analytical Queries -** Jinyu & Yichen: will write and test at least 8 analytical queries for the insight.
- **Web Front End -** Yichen: Build a user interface to display the dashboard and query results.

- **Chatbot integration -** Richard: develop a natural language interface for ranking questions.
- **Deployment in AWS -** Yichen & Richard: Deploy our web on AWS using CloudFront and S3.

2.3 Missing Skills

- We have limited experience with NLP and Sentiment Analysis. We need a short period of time to learn and understand the basics, so that we can fine-tune the model properly.
- Chatbot logic may need additional learning on SQL Agent frameworks.

2.4 Existing Experience and Must-Learn

- **Already Experienced:** We are familiar with Python, SQL, JavaScript as programming languages, and MySQL databases. We've also learned the data processing tools like pandas and numpy. We are experienced with the basic web development tools like HTML, CSS, React, and Flask. We also have some basic knowledge of version control (GitHub)
- **To Be Learned:** We must explore the cloud deployment on AWS, including using EC2, S3, and RDS console to build our solution on the cloud. We also need to learn some basics of NLP, as we need to fine-tune the model and perform sentiment analysis.

2.5 Programming Languages and Platform

- **Programming Language:** Python, SQL, and React
- **Deployment** AWS

3 Skills and Tools Assessment

3.1 External Resources and Assistance

- **Kaggle and Hugging Face tutorials** for model fine-tuning, text tokenization, and sentiment labeling analysis.
- **AWS educational documentation** for developing our web using EC2 and S3, if necessary.

3.2 Tools, Frameworks, and Libraries

- **Database:** MySQL / PostgreSQL for structured data storage and analytical SQL queries(products, users, reviews, etc.)
- **Data Cleaning & Preprocessing:** Python and Pandas for data cleaning, null handling, and normalization
- **NLP Processing** NLTK or TextBlob for tokenization and sentiment analysis
- **Frontend:** React with HTML/CSS/JavaScript for user dashboards
- **Cloud:** AWS EC2, S3, and RDS for storing data in the cloud

3.3 Ensuring Proficiency and Comfort

- Early Discussion: Every team member discussed together, and we all chose the part we are familiar with.
- Shared Documentation: We will maintain a GitHub repository with a detailed README, guides, and environment settings.
- Weekly Meeting: We will have a short meeting each week to demo progress.

3.4 Task Allocation and Role Clarity

- **Jinyu Li** - Data preprocessing, SQL database design, and analytical query implementation.
- **Yichen Zhang** – Front-end development, user interface design, and SQL query integration on the web dashboard.
- **Richard Chen** – Backend API development, NLP model fine-tuning, and chatbot integration.

4 Initial Setup

4.1 Development Environment

- **Programming Languages:** Python(3.11 or above), SQL, and JavaScript/React.
- **Database:** MySQL stalled in the local for now
- **Backend Framework:** Flask or FastAPI
- **Frontend Framework:** React using Node.js and npm for dependency management
- **Cloud** AWS EC2, S3, and RDS instances.

4.2 Version Control and Repository Access

We all have access to the public repository using GitHub.