Healthcare Analytics Pipeline Project for MediCare Hospital

Introduction

MediCare Hospital invites proposals for a six-week data engineering capstone project focusing on healthcare data. The healthcare industry generates massive volumes of patient and treatment data daily, making efficient data pipelines critical for improving outcomes. Teams will assume the role of data engineers in a hospital analytics department, tasked with cleaning, integrating, and analyzing patient records and health metrics. Emphasis will be on maintaining data quality and compliance in a sensitive domain.

Project Scope and Objectives

This project simulates real-world healthcare data challenges. Teams will:

- **Data Acquisition:** Identify and integrate publicly available healthcare datasets such as Electronic Health Records (EHR) samples or hospital admissions data (e.g., a sanitized subset of the UCI "Diabetes 130-US Hospitals" dataset or Kaggle's COVID-19 patient statistics). Ingest data via CSV files or APIs, ensuring that any patient identifiers are anonymized or omitted.

- **Data Cleaning:** Address missing clinical values (e.g., absent lab results), correct data entry errors, and standardize formats (ICD codes, measurement units). Document imputation and cleaning strategies.

- **Exploratory Analysis:** Perform EDA to uncover health trends (e.g., disease incidence by age or location, hospital admission rates). Create visualizations such as demographic heatmaps, time-series of admissions, or distributions of vital signs.

- **Data Modeling:** Implement analytical tasks, for example predictive modeling (e.g., patient readmission risk or length-of-stay prediction) or clustering (e.g., segment patients by condition). Ensure models account for healthcare context and are validated properly.

- **Data Storage & Governance:** Propose a secure storage solution (HIPAA-compliant data warehouse or secure database) with data validation checks. Optimize for query performance while ensuring confidentiality and data

integrity.

Teams must handle data responsibly: protect patient privacy at all stages and follow ethical guidelines. Working with realistic healthcare datasets will expose students to challenges like data sensitivity and regulatory constraints, which are key in healthcare informatics.

Phases and Weekly Deliverables

The project is divided into weekly phases:

- **Phase 1 (Weeks 1–2): Data Collection and Initial Processing**
    - *Week 1:* Source relevant healthcare datasets (e.g., hospital admission records, patient surveys, sensor data). Gain understanding of data schemas and privacy requirements. **Deliverable:** Data source catalog and schema overview.
    - *Week 2:* Clean and preprocess data: handle missing values with appropriate medical imputation techniques, anonymize or remove sensitive identifiers, and standardize units. **Deliverable:** Cleaned dataset and data dictionary with lineage documentation.
- **Phase 2 (Weeks 3–4): Data Integration and Transformation**
    - *Week 3:* Build robust data pipelines to load cleaned data into a secure database or data warehouse (e.g., SQL or cloud database). Automate ETL jobs (for example, using Python scripts scheduled via Airflow). **Deliverable:** Ingestion pipeline scripts and demonstration of automated data loading.
    - *Week 4:* Transform data for analysis: aggregate patient outcomes, compute derived health metrics (e.g., BMI from weight/height), and partition data by department or timeframe. Implement database optimizations (indexing, partitioning) to speed up analytical queries. **Deliverable:** Transformed data tables and performance benchmark report.
- **Phase 3 (Weeks 5–6): Analysis, Visualization, and Reporting**

- *Week 5:* Conduct advanced analysis: for instance, predict patient readmission risk or identify factors affecting treatment success. Validate and interpret model results with domain experts. **Deliverable:** Modeling code and evaluation report.

- *Week 6:* Develop visual dashboards or summary reports for stakeholders (e.g., interactive charts of patient outcomes, anomaly detection alerts). Prepare the final presentation of findings and recommendations. **Deliverable:** Interactive visualization dashboard and final project presentation.

Weekly mentor meetings will ensure compliance with healthcare data standards. Teams should split tasks (e.g., pipeline engineering, modeling, reporting) but all members should understand each step. Collaboration and peer code review are expected.

Use of Generative AI

Generative AI tools should be used thoughtfully to augment the workflow:

- **ChatGPT (OpenAI):** Use ChatGPT to generate code templates (e.g., for sanitization scripts or SQL queries) and to assist with healthcare data questions (such as handling clinical codes or statistical tests). ChatGPT can also help summarize technical concepts or propose imputation strategies.

- **GitHub Copilot:** Leverage Copilot within the IDE to speed up writing ETL pipelines and SQL queries, such as providing autocompletion for data transformation loops or database queries.

- **Gemini (Google):** Utilize Gemini for researching medical data standards and for complex code suggestions. For example, Gemini can help find libraries for healthcare data processing or suggest efficient database designs based on domain needs.

- **Claude (Anthropic):** Apply Claude for summarizing technical documentation or patient case descriptions in clear language. Claude's focus on safety and clarity makes it useful for generating reliable explanations of model results or writing sections of reports.

- **LangChain:** Employ LangChain to build chains of prompts that facilitate analysis. For example, use LangChain to feed summarized patient data to an LLM and

retrieve natural language summaries of key trends or to sequence prompts for iterative data quality checks.

Students should document how each AI tool was used, highlighting improvements in efficiency or insights. For instance, note if ChatGPT helped draft a data validation routine or if Claude provided a clearer explanation of a complex chart.

Expected Outcomes

Teams completing this project will:

- Demonstrate secure ingestion and storage of sensitive health data, with privacy-preserving transformations and validations.
- Clean and integrate multi-source medical datasets into a unified analytics repository.
- Perform exploratory and predictive analysis on healthcare outcomes (e.g., factors influencing readmission risk).
- Visualize health metrics effectively (e.g., patient outcome dashboards, anomaly detection plots).
- Integrate generative AI tools to streamline coding and reporting, reflecting on the benefits and limitations of these tools in a healthcare context.

This hands-on experience will prepare students for roles in health data engineering, emphasizing both technical skills and ethical data practices.

Deliverables

- A secured repository (e.g., private GitHub) containing all pipeline code, transformation scripts, and model code, with thorough documentation.
- Cleaned healthcare datasets along with detailed data dictionaries explaining each field and its transformation.

- Written reports for each phase documenting methods, challenges (such as privacy concerns), and results.
- Final analytical report summarizing predictive model outcomes and insights on patient data trends.
- Visual dashboards or reports (using tools like Tableau, Power BI, or Python libraries) showing healthcare insights (e.g., admission trends, outcome distributions).
- A final presentation (slides or recorded demo) summarizing the approach, findings, and how AI tools assisted the project.

Proposal Submission Requirements

Submit a PDF proposal including:

- **Problem Understanding:** Brief overview of the healthcare question being addressed and project goals.
- **Data Strategy:** Description of chosen dataset(s), how privacy/anonymization will be handled, and data access methods.
- **Technical Plan:** Tools and technologies to be used (databases, languages, frameworks), and high-level architecture.
- **Timeline:** Week-by-week plan (1–6) with assigned responsibilities.
- **Team Profile:** Background of team members, including any relevant experience (e.g., prior data projects, healthcare domain knowledge).

Proposals should emphasize ethical data handling (no real patient data) and effective use of public datasets and AI tools.

Evaluation Criteria

Submissions and projects will be evaluated on:

1. **Data Privacy & Security:** Adequacy of measures to protect patient privacy and data security.

2. **Technical Execution:** Efficiency and robustness of data pipelines and analysis methods.

3. **Insight and Impact:** Relevance and depth of healthcare insights derived (e.g., actionable findings for patient care).

4. **AI Integration:** Thoughtful integration of AI tools (e.g., using ChatGPT to assist rather than blindly trust outputs).

5. **Communication:** Clarity and professionalism of documentation, reports, and visualizations in a healthcare context.

Submission Deadline

The Project is due by **July 10, 2025 (Tentative)** via email.