City-Health-Insights-Proposal

Project Proposal: City Health Insights - Centralized Patient Data Pipeline

1. Project Description

The project aims to build a centralized healthcare data pipeline and machine learning analytics platform for city-wide healthcare data integration. It collects, processes, and analyzes patient, doctor, and symptom data from various healthcare providers to enhance healthcare delivery, improve patient outcomes, and optimize operational efficiency. Leveraging Azure Health Data Services, SQL databases, Apache Airflow orchestration, and Power BI dashboards, the system creates an integrated healthcare ecosystem with advanced analytics and real-time capabilities.

2. Group Members & Roles

Member 1 (Ahmed Adel)

- Data Pipeline Orchestrator (Airflow + Azure + Architecture)

- Lead project planning, architecture design, and documentation aligned with Azure Health Data Services framework.

- Manage Azure cloud environment, CI/CD pipelines, and orchestrate data workflows using Apache Airflow.

- Coordinate communication and milestone delivery across data engineering, ML, and visualization teams.

Member 2 (Mohamed Nasser)

– Data Preprocessing & Machine Learning Engineer (Python + Pandas + PySpark + Scikit-learn)

- Responsible for data collection, cleaning, and feature engineering following healthcare data standards.

- Develop and train machine learning models for patient risk prediction and symptom-based doctor recommendations.

- Continuously evaluate model metrics (accuracy, precision, recall) and deploy models integrated via Airflow or dbt pipelines.

Member 3 (**Maryam Ahmed**)

– Database & Transformation Engineer (SQL + dbt + Hive / Azure SQL)

- Design normalized database schemas compliant with FHIR and HL7 standards for patient and medical data.

- Build dbt models to transform raw data through medallion layers (Bronze → Silver → Gold).

- Optimize SQL queries to ensure data consistency and high query accuracy (≥ 95%) for analytics and reporting.

Member 4 (Noran Salem)

– Big Data & Streaming Engineer (Kafka + Spark + Hadoop)

- Implement streaming infrastructure for real-time ingestion of medical device data and patient vitals using Kafka and Spark Structured Streaming.

- Manage HDFS or Hive storage optimization, data partitioning, and integration of streaming with batch data pipelines for hybrid workflows.

Member 5 (Walaa Mohamed)

–Visualization & Reporting Specialist (Power BI + Azure + Reporting)

- Develop interactive Power BI dashboards visualizing patient trends, hospital resource utilization, and ML insights.

- Optimize dashboard performance ensuring load times under 3 seconds.

- Prepare and deliver final presentation slides and comprehensive project reports.

- Collect stakeholder feedback and ensure completeness of deliverables (target: 100%).

3. Team Leader

- Maryam Ahmed Gamal Eldin (Mariaamahmed489@gmail.com)

4. Objectives

- Establish a unified data pipeline aggregating healthcare data from hospitals, clinics, labs, and IoT medical devices across the city.
- Implement advanced ML analytics to predict patient risks, optimize treatment paths, and provide personalized doctor recommendations.
- Create comprehensive and responsive user interfaces and dashboards for patients, doctors, and healthcare administrators.
- Enable real-time critical alerts and advanced population health management analytics.
- Improve healthcare operational efficiency via data-driven resource allocation and service quality monitoring.

5. Tools & Technologies

- Azure Health Data Services leveraging FHIR and DICOM standards for secure, compliant healthcare data management.
- Apache Airflow for robust orchestration of data pipelines and ML model deployment.
- Python libraries (Pandas, PySpark, Scikit-learn) for data preprocessing and machine learning.
- SQL and dbt for database schema design and layered data transformations.
- Kafka, Spark, and Hadoop for big data streaming and batch processing hybrid pipelines.
- Power BI for interactive visual analytics and reporting.
- Git & CI/CD for version control and automated deployment.

6. Milestones & Deadlines

Phase	Tasks	Deadline
Phase 1: Foundation & Infra	Azure infrastructure setup, user authentication, initial ETL and data ingestion pipelines.	12/10 : 22/10
Phase 2: Core Development	Build user interfaces, basic dashboards, appointment management; deploy initial ML models.	23/10 : 3/11
Phase 3: Advanced Analytics	Integrate predictive ML models, real-time alerts, epidemic surveillance, and population health tools.	4/11 : 14/11
Phase 4: Integration & Expansion	Complete full city-wide facility integration, telemedicine platform, blockchain security.	21/11 : 28/11
Phase 5: Optimization & Launch	Optimize system based on pilot feedback, document processes, launch city-wide deployment.	29/11 : 6/12

7. KPIs (Key Performance Indicators)

Data Preprocessing

- 100% of missing and duplicate data handled accurately following healthcare data standards.
- Data preprocessing scripts execution occurring within expected performance thresholds.

SQL Integration

- SQL query accuracy meeting or exceeding 95% against expected results.
- Average query execution time maintained below performance targets for test queries.

Visualization

- Dashboard load time maintained under 3 seconds for optimal user experience.

- Visualization coverage of at least 90% of required KPIs and metrics across dashboards.

Presentation

- Complete project documentation and reporting with 100% of required sections delivered.
- Achieve stakeholder clarity and satisfaction score of at least 4 out of 5.