****

Data warehousing(dwh)

**final report**

**CEP**



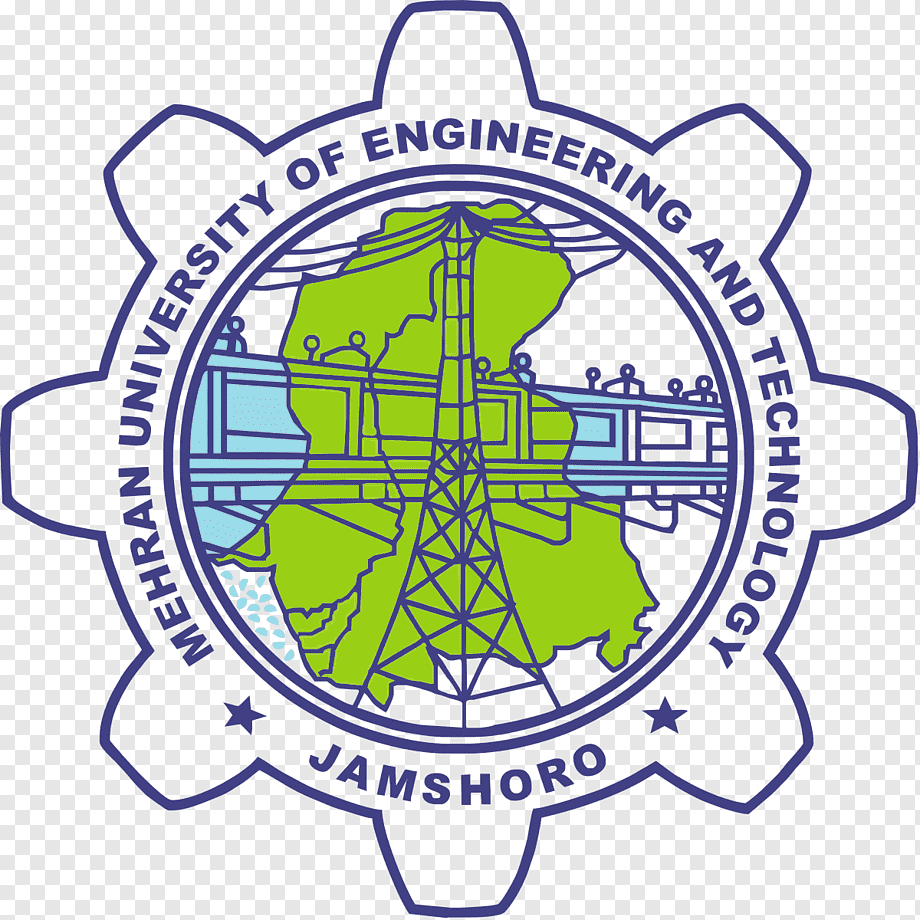
**Submitted to: Dr. Naeem**

**Submitted by: 23SW050 (Ghulam Hussain)**

**23SW056 (Abbas Ali)**

April 15, 2025

MUET JamshOro

******

**1. Introduction**

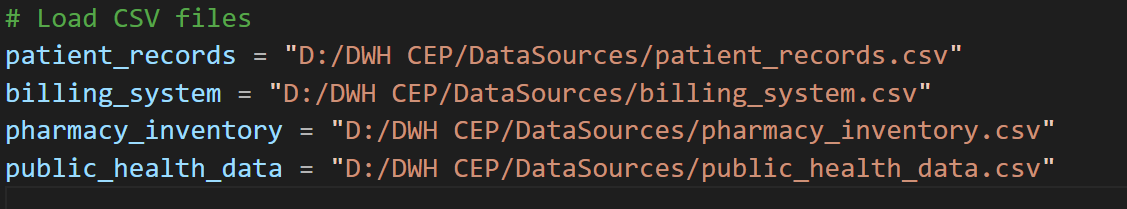
This report presents the design and implementation of a healthcare data warehouse aimed at enhancing patient care, optimizing resource allocation, and improving operational efficiency. The project integrates data from multiple sources, including patient records, billing systems, pharmacy inventory, and public health statistics, to support analytical queries and informed decision-making.​

**2. ETL Process Design**

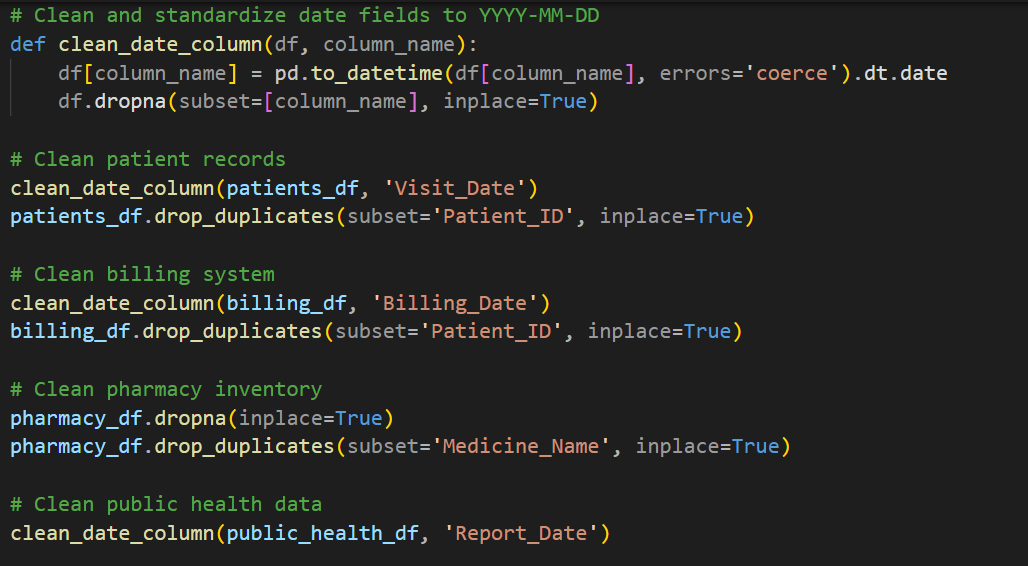
The Extract, Transform, Load (ETL) process was developed using Python with Pandas and SQLAlchemy to handle data extraction from CSV files, transformation for consistency, and loading into a MySQL database.​

**Key Steps:**

* **Extraction:** Data was extracted from four primary sources:​
  + Patient Records
  + Billing System​
  + Pharmacy Inventory​
  + Public Health Data​



* **Transformation:** Data inconsistencies were addressed by:​
  + Standardizing date formats to YYYY-MM-DD
  + Handling missing values by dropping incomplete records where necessary.​
  + Removing duplicate entries to ensure data integrity.​



* **Loading:** Transformed data was loaded into corresponding dimension and fact tables in the data warehouse.​

**Data Quality Issues Addressed:**

1. **Inconsistent Date Formats:** Resolved by converting all date fields to a standard format using pd.to\_datetime().​
2. **Missing Values:** Handled by dropping records with critical missing information to maintain data quality.​

**3. Dimensional Modeling**

A star schema was implemented to facilitate efficient querying and reporting.​

**Dimension Tables:**

* **Patient\_Dim:** Stores patient demographics.​
* **Diagnosis\_Dim:** Contains diagnosis codes and descriptions.​
* **Date\_Dim:** Captures date-related information for analysis over time.​
* **Medication\_Dim:** Holds details about medications.​
* **Supplier\_Dim:** Information about medication suppliers.​
* **ZipCode\_Dim:** Geographical data for regional analysis.​

**Fact Tables:**

* **Fact\_TreatmentCosts:** Records costs associated with patient treatments.​
* **Fact\_MedicationDemand:** Tracks medication dispensing and stock levels.​
* **Fact\_OutbreakCorrelation:** Links public health data with hospital visit statistics.​

**Slowly Changing Dimension:**

The Patient\_Dim table is designed to handle slowly changing dimensions, such as changes in patient age over time, ensuring historical accuracy in reporting.​

**Conceptual Design:**

A diagram of a medical procedure

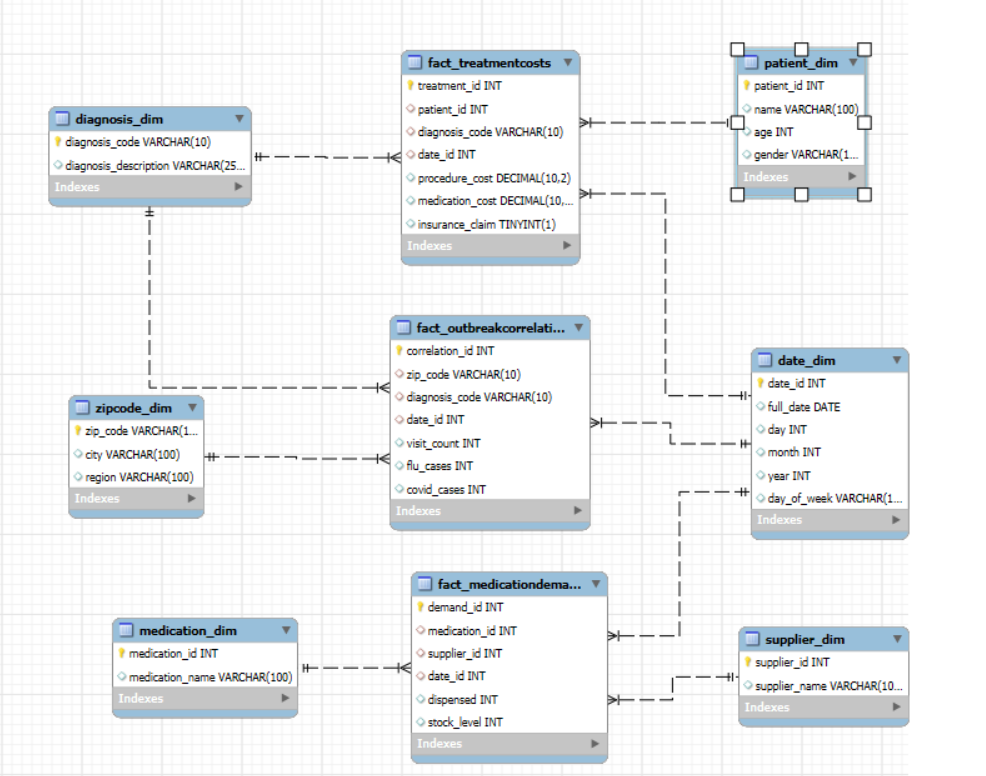
AI-generated content may be incorrect.

**Logical Design:**

A diagram of a data flow

AI-generated content may be incorrect.

**Physical Design:**



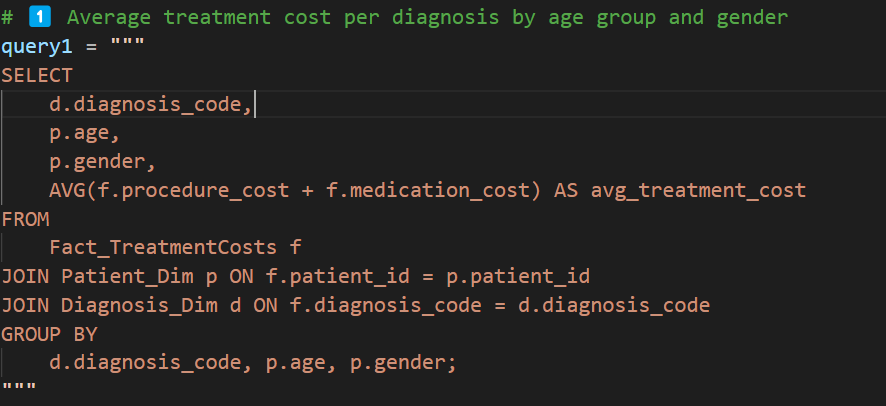
Queries file is attached to it

**4. OLAP Queries and Analysis**

Online Analytical Processing (OLAP) queries were developed to support the following analyses:​

1. **Average Treatment Cost per Diagnosis by Age Group and Gender:**

This query calculates the average cost of treatments, segmented by diagnosis, age group, and gender, aiding in cost analysis and budgeting.



**Result:**

A screenshot of a computer

AI-generated content may be incorrect.

1. **Medication Stock Levels and Replenishment Alerts:**

Analyzes current stock levels against dispensing rates to generate alerts for replenishment, ensuring medication availability.

A computer screen with text and images

AI-generated content may be incorrect.

**Result:**

A screenshot of a computer

AI-generated content may be incorrect.

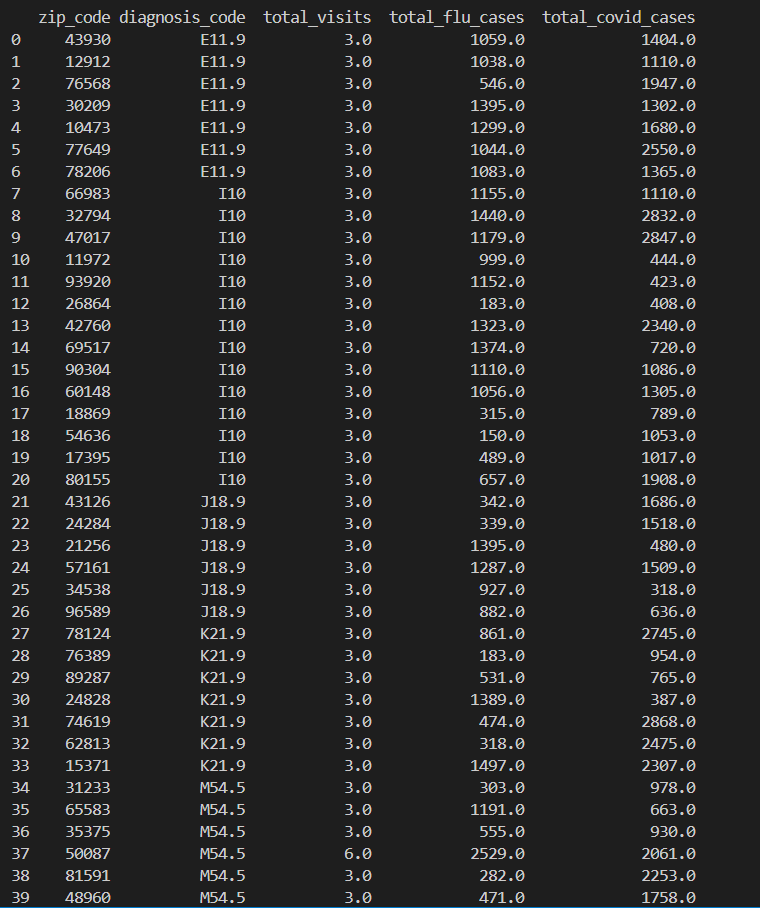
1. **Correlation Between Disease Outbreaks and Hospital Visit Spikes:**

Examines the relationship between public health data on disease outbreaks and corresponding increases in hospital visits, supporting proactive healthcare responses.

A computer screen with text

AI-generated content may be incorrect.

**Result:**

****

**5. Performance Optimization**

To enhance query performance and ensure scalability:

* **Indexing:** Indexes were created on foreign key columns to speed up join operations.​
* **Partitioning:** Fact tables were partitioned by date to improve query efficiency for time-based analyses.​

**6. Scalability Considerations**

Anticipating a tenfold increase in data volume over five years due to the expansion to 50 hospitals, the following design modifications are proposed:​

* **Modular Architecture:** Implementing a modular data warehouse structure to allow independent scaling of components.​
* **Cloud Integration:** Transitioning to a cloud-based data warehouse solution to leverage scalable storage and compute resources.​

**7. Research Findings**

**Data Warehouse Architectures:**

* **Enterprise Data Warehouse (EDW):** Centralized repository for all organizational data, facilitating comprehensive analysis.​
* **Data Mart:** Focused subsets of data tailored to specific departments or functions.​[Medium](https://medium.com/%40IntelliSoft/data-warehousing-in-healthcare-transformative-strategies-23a75db45170?utm_source=chatgpt.com)
* **Operational Data Store (ODS):** Real-time data integration for operational reporting.​

**Physical Schemas:**

* **Star Schema:** Simplified design with a central fact table connected to dimension tables, enabling fast query performance.​
* **Snowflake Schema:** Normalized dimension tables to reduce data redundancy, at the cost of more complex queries.​

**Pros and Cons:**

* **Star Schema:**
  + *Pros:* Simplified queries, improved performance.​
  + *Cons:* Potential data redundancy.​
* **Snowflake Schema:**
  + *Pros:* Reduced data redundancy, better data integrity.​
  + *Cons:* More complex queries, potentially slower performance.​

**8. Assumptions**

* Data sources are reliable and updated regularly.​
* Users require access to both current and historical data for analysis.​
* Security and privacy regulations are adhered to in data handling.​

**9. Conclusion**

The implemented healthcare data warehouse successfully integrates diverse data sources to provide actionable insights into patient care, resource management, and operational efficiency. The design accommodates current analytical needs and is scalable to meet future demands as the healthcare provider expands.​