Data Warehouse Implementation Report

1. Assignment Overview

This report presents the **design and implementation of a data warehouse** using **Python for data storage and Snowflake for ETL and analytics**. The project consists of **sourcing raw data, transforming it, and structuring it into a star schema** to facilitate analytical querying and business intelligence insights.

2. Data Sources

The dataset chosen for the assignment was sourced from the **data.CMS.gov**, which provides a detailed overview of Hospital Operations and Patients Data from multiple sources. A total of **100,000 records** are used from the raw data, covering aspects such as facilty ID, facilty name, measure name, address(state, zipcode, etc.) among others.

Dataset Overview

Multiple patient-related datasets form the basis of the hospital data warehouse for collecting quality-driven healthcare information and performance metrics. Various facilities operating in the United States can find structured healthcare data regarding hospitals combined with quality measurements, payment details, and infection statistics in this dataset.

Potential Use Cases:

This dataset can be leveraged for various healthcare and business intelligence applications. Some of the potential use cases include:

1. Hospital Performance Benchmarking
    - Compare hospital quality scores against national benchmarks.
    - Identify hospitals with the highest and lowest patient outcomes.
2. Healthcare Cost Analysis
    - Analyze hospital payment trends based on ownership, type, and geographical location.
    - Evaluate financial efficiency across different healthcare providers.
3. Identifying High-Risk Hospitals
    - Detect hospitals with consistently poor scores in complications, infections, and mortality rates.
    - Provide insights for government and insurance bodies to improve healthcare standards.
4. Impact of Ownership and Hospital Type on Care Quality

- Analyze differences in patient care between for-profit, non-profit, and government hospitals.
- Study whether hospital type (academic medical centers, rural hospitals, specialty hospitals) affects quality scores.

5. Predictive Healthcare Modeling
    - Use machine learning techniques to predict hospital performance based on historical trends.
    - Identify key factors contributing to high or low hospital ratings.
6. Public Health Policy and Decision Making
    - Provide data-driven insights to health policymakers for targeted intervention.
    - Help in fund allocation to underperforming hospitals to improve care standards.

Dataset Files

The following datasets were used:

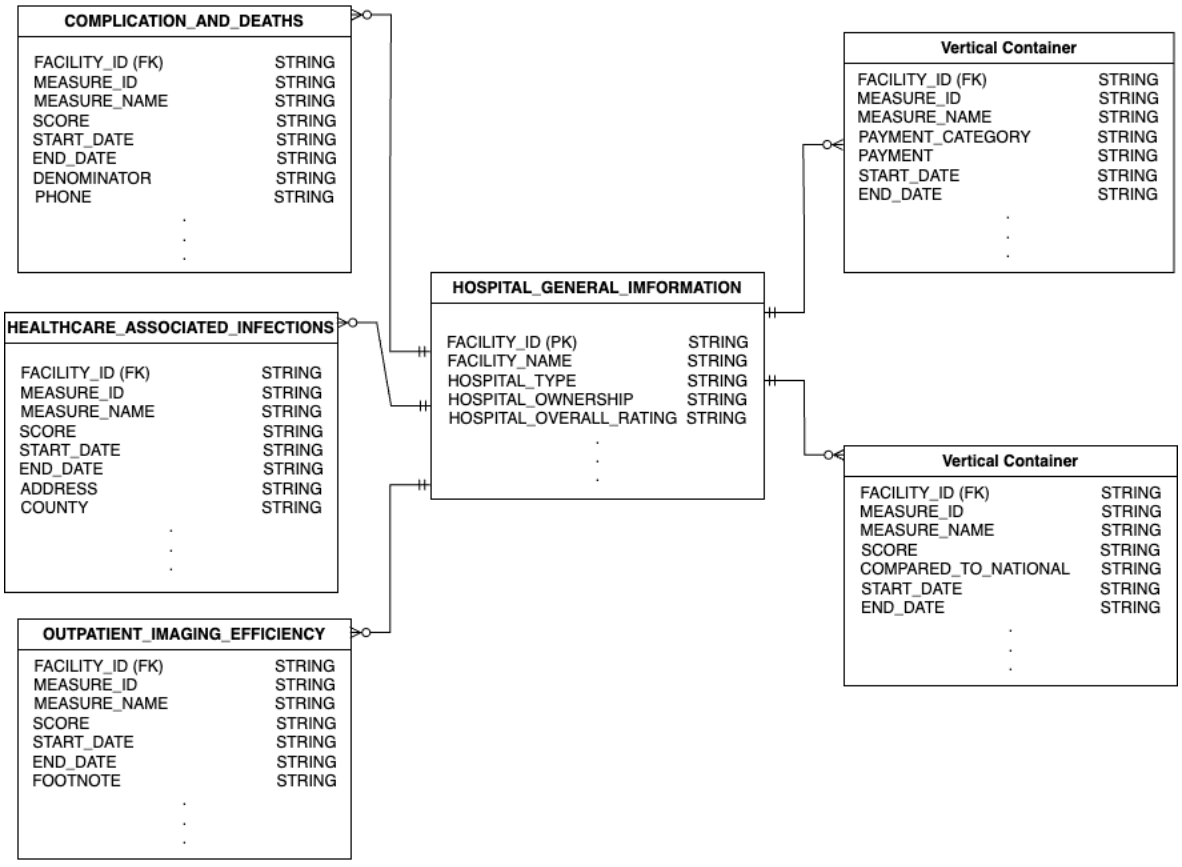
Dataset	File Name	Description
Hospital General Information	hospital_general_information.*csv*	Contains details about each hospital, such as name, type, ownership, emergency services, and ratings.
Complications and Deaths	complications_and_deaths.*csv*	Tracks complications and death rates for various procedures and conditions.
Outpatient Imaging Efficiency	outpatient_imaging_efficiency.*csv*	Contains efficiency ratings for outpatient imaging services.
Healthcare Associated Infections	Healthcare_associated_infections.*csv*	Includes infection rates for hospitals.
Timely Effective Care	Timely_effective_care.*csv*	Measures timely care effectiveness.
Payment and Value of Care	payment_and_value_of_care.*csv*	Contains hospital payments and cost-related measures.

The full data dictionary, describing all fields in the dataset, is provided separately in the **README file** accompanying this report.

3. Normalised Database

The normalized hospital data organizes healthcare metrics into five distinct tables that all relate to the central hospital_general_information table. This structure eliminates redundancy by storing facility details (like name, location, and ownership type) only once in the main table, while specific performance metrics are separated into specialized tables: complications_and_deaths tracks mortality outcomes, healthcare_associated_infections monitors infection rates, outpatient_imaging_efficiency measures diagnostic performance, payment_and_value_of_care contains cost metrics, and timely_and_effective_care focuses on treatment timeliness. Each of these tables connects back to the hospital_general_information table through the Facility_ID foreign key, creating a clean one-to-many relationship model. This normalized approach allows for efficient data storage, reduces update anomalies, and enables more flexible querying across different measurement domains while maintaining data integrity.

ER Diagram



4. ETL Process Implementation

The ETL process follow the steps below:

- Extract: Data is loaded from the raw CSV file.
- Transform: Data is cleaned, standardized, and normalized.
- Load: Normalized data is inserted into staging tables, then transformed into dimensions and fact tables.

ETL Challenges and Solutions:

- Handling Missing Data: NULL values replaced with defaults where applicable.
- Surrogate Keys: Used for dimensions to maintain consistency.
- Slowly Changing Dimensions: Implemented tracking for changes over time.

ETL Execution Logs:

Screenshots and logs of successful execution are included in the README file.

5. Data Warehouse Design: Star Schema

5.1 Dimension Tables:

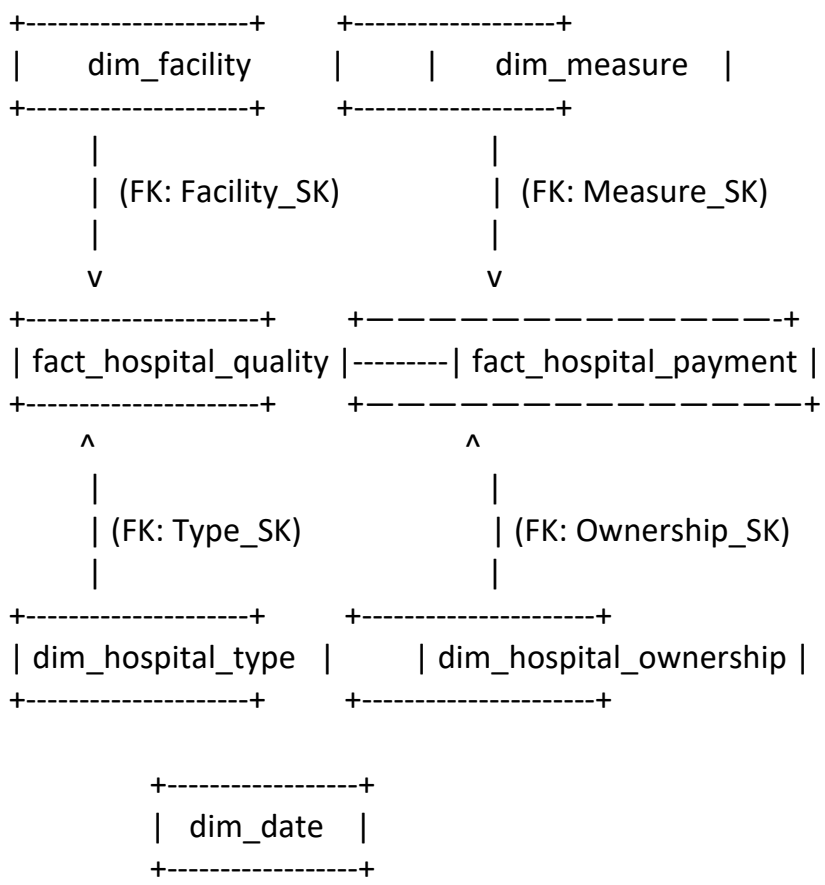
Table Name	Description
dim_facility	Contains hospital facility attributes like name, address, type, and ownership.
dim_measure	Stores the various healthcare quality measures being tracked.
dim_date	Holds date-related attributes for time analysis.
dim_hospital_type	Contains the different categories of hospital types.
dim_hospital_ownership	Lists the various ownership models for hospitals.

5.2 Fact Tables:

Table Name	Description
Fact_hospital_quality	Records quality measurements like scores and comparisons to national benchmarks.
fact_hospital_payment	Stores payment data including amounts, estimates, and payment categories.

5.3 Star Schema Design

Visual Representation of the Star Schema:



1. Fact Table: fact_hospital_quality
Primary Fact Table – Stores hospital quality-related performance metrics. Connected to:

- dim_facility → To track which facility the quality measure belongs to (Facility_SK).
- dim_measure → To identify the specific quality measure being evaluated (Measure_SK).
- dim_date → To track when the quality measure was recorded (Date_SK).
- dim_hospital_type → To categorize hospitals based on their type (Type_SK).
- Dim_hospital_ownership → To identify hospital ownership details (Ownership_SK).

2. Fact Table: fact_hospital_payment
Stores financial data related to hospital payments and value-based care. Connected to:

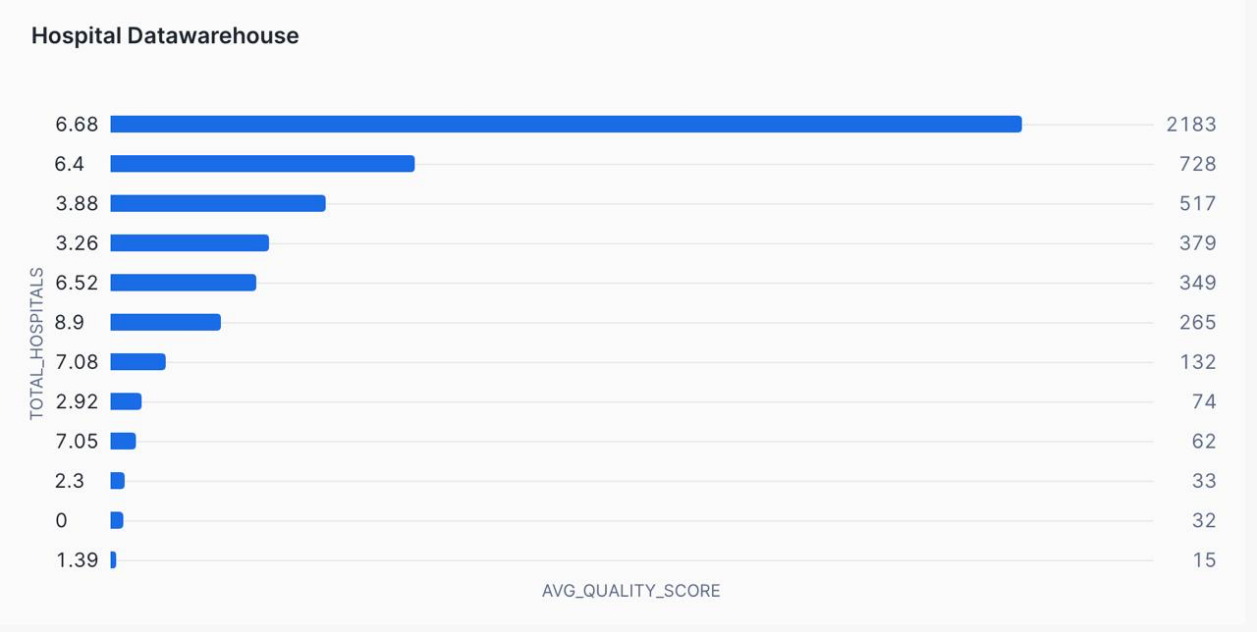
- dim_facility → To track which facility the payment data belongs to (Facility_SK).
- dim_measure → To identify the specific payment-related measure (Measure_SK).
- dim_date → To track when the payment was recorded (Date_SK).
- dim_hospital_ownership → To classify facilities based on their ownership type (Ownership_SK).

6. Analytical Querying and Business Insights

6.1 Query 1: Identify which hospital ownership types have the best and worst quality scores.

Business Insight: Which ownership types consistently perform best in terms of quality?

```
SELECT
    o.Hospital_Ownership,
    COUNT(DISTINCT f.Facility_SK) AS Total_Hospitals,
    ROUND(AVG(q.Score), 2) AS Avg_Quality_Score,
    SUM(CASE WHEN q.Compared_to_National = 'Worse' THEN 1 ELSE 0 END) AS
Worse_Than_National,
    SUM(CASE WHEN q.Compared_to_National = 'Better' THEN 1 ELSE 0 END) AS
Better_Than_National,
    (SUM(CASE WHEN q.Compared_to_National = 'Better' THEN 1 ELSE 0 END) * 100.0) /
    NULLIF(COUNT(q.Quality_Fact_SK), 0) AS Percent_Better
FROM dimensional.fact_hospital_quality q
JOIN dimensional.dim_facility f ON q.Facility_SK = f.Facility_SK
JOIN dimensional.dim_hospital_ownership o ON f.Hospital_Ownership = o.Hospital_Ownership
JOIN dimensional.dim_measure m ON q.Measure_SK = m.Measure_SK
WHERE q.Fact_Type = 'Complication' -- Focus on complications
GROUP BY o.Hospital_Ownership
ORDER BY Percent_Better DESC;
```



6.2 Query 2: To analyze how hospital ownership, type, quality scores, and payments impact performance and financial efficiency.

Business Insight: This analysis helps identify disparities in hospital performance across ownership types and categories, guiding policy interventions and resource allocation

```
SELECT
    o.Hospital_Ownership,
    t.Hospital_Type,
    COUNT(DISTINCT f.Facility_SK) AS Total_Hospitals,
    ROUND(AVG(q.Score), 2) AS Avg_Quality_Score,
    ROUND(AVG(p.Payment), 2) AS Avg_Payment,
    SUM(CASE WHEN q.Compared_to_National = 'Worse' THEN 1 ELSE 0 END) AS
Worse_Than_National,
    SUM(CASE WHEN q.Compared_to_National = 'Better' THEN 1 ELSE 0 END) AS
Better_Than_National,
    ROUND((SUM(CASE WHEN q.Compared_to_National = 'Better' THEN 1 ELSE 0 END) * 100.0) /
NULLIF(COUNT(q.Quality_Fact_SK), 0), 2) AS Percent_Better,
    ROUND((SUM(CASE WHEN q.Compared_to_National = 'Worse' THEN 1 ELSE 0 END) * 100.0) /
NULLIF(COUNT(q.Quality_Fact_SK), 0), 2) AS Percent_Worse
FROM dimensional.fact_hospital_quality q
JOIN dimensional.fact_hospital_payment p ON q.Facility_SK = p.Facility_SK
JOIN dimensional.dim_facility f ON q.Facility_SK = f.Facility_SK
JOIN dimensional.dim_hospital_ownership o ON f.Hospital_Ownership = o.Hospital_Ownership
JOIN dimensional.dim_hospital_type t ON f.Hospital_Type = t.Hospital_Type
WHERE q.Fact_Type = 'Complication' -- Focus on complications
GROUP BY o.Hospital_Ownership, t.Hospital_Type
ORDER BY Percent_Worse DESC, Avg_Quality_Score ASC;
```

6.3 Query 3: Identifies states with high infection rates and contributing hospital types.

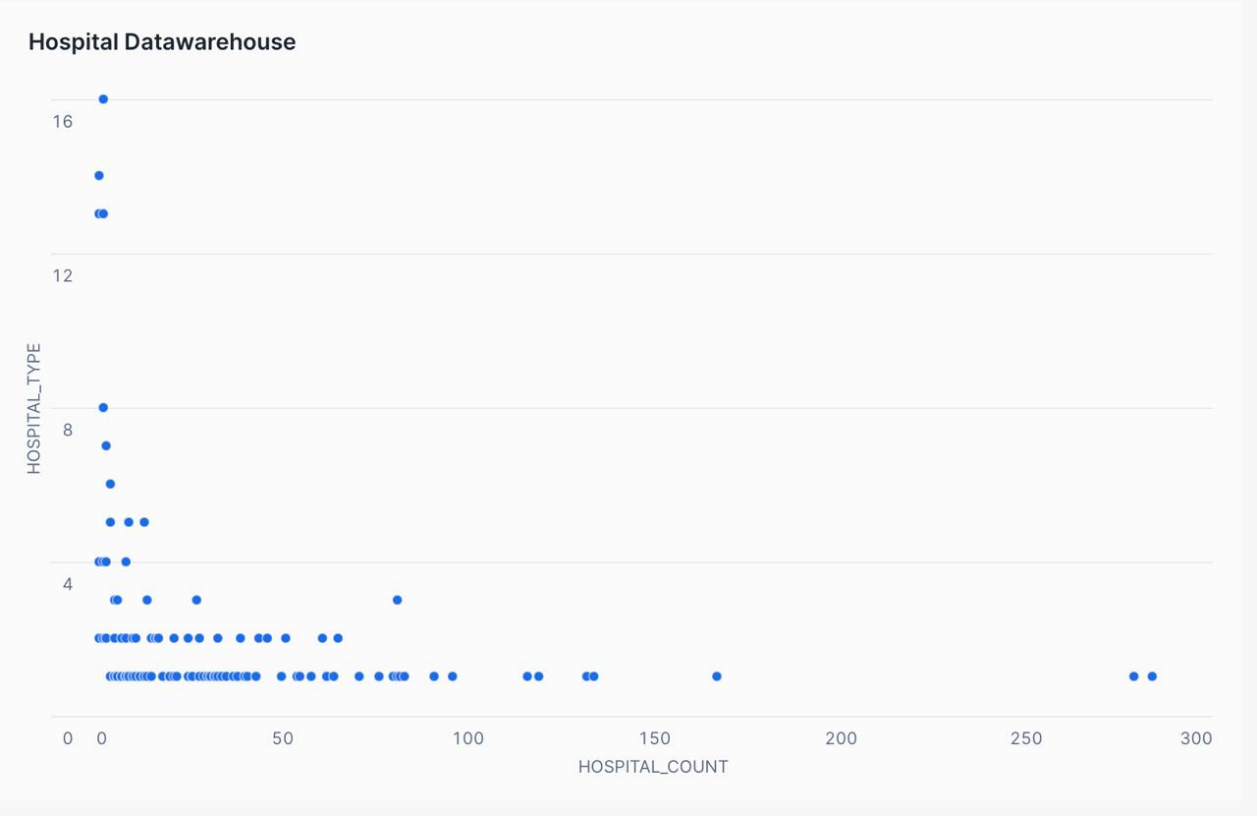
Business Insight: Identifies states with high infection rates, guiding policymakers and hospitals in improving healthcare safety and infection control.

```
WITH InfectionRanks AS ( SELECT
        f.Facility_Name,
        f.State,
        t.Hospital_Type,
        i.Score AS Infection_Rate,
        q.Score AS Quality_Score,
        RANK() OVER (PARTITION BY f.State ORDER BY i.Score DESC) AS Infection_Rank,
        RANK() OVER (PARTITION BY f.State ORDER BY q.Score ASC) AS Quality_Rank
    FROM dimensional.fact_hospital_quality q
    JOIN dimensional.fact_hospital_quality i ON q.Facility_SK = i.Facility_SK
    JOIN dimensional.dim_facility f ON q.Facility_SK = f.Facility_SK
    JOIN dimensional.dim_hospital_type t ON f.Hospital_Type = t.Hospital_Type
    WHERE i.Score IS NOT NULL AND q.Score IS NOT NULL),
WorstHospitals AS (
    SELECT
        Facility_Name, State, Hospital_Type, Infection_Rate, Quality_Score
    FROM InfectionRanks
    WHERE Infection_Rank <= 10 AND Quality_Rank <= 10  -- Worst 10 hospitals per
state)
SELECT
    State,
    Hospital_Type,
    COUNT(*) AS Num_Hospitals,
    ROUND(AVG(Infection_Rate), 2) AS Avg_Infection_Rate,
    ROUND(AVG(Quality_Score), 2) AS Avg_Quality_Score
FROM WorstHospitals
GROUP BY State, Hospital_Type
ORDER BY Avg_Infection_Rate DESC, Avg_Quality_Score ASC;
```

6.4 Query 4: To analyze the distribution of hospital types across states, identifying variations in healthcare infrastructure and availability.

Business Insight: This analysis helps identify regional disparities in hospital types, guiding resource allocation and healthcare infrastructure planning.

```
SELECT
    f.State,
    t.Hospital_Type,
    COUNT(f.Facility_SK) AS Hospital_Count
FROM dimensional.dim_facility f
JOIN dimensional.dim_hospital_type t ON f.Hospital_Type = t.Hospital_Type
GROUP BY f.State, t.Hospital_Type
ORDER BY f.State, Hospital_Count DESC;
```



7. Conclusion and Future Improvements

The implemented hospital data warehouse project efficiently merges different health care datasets into an organized system that supports analysis. We created a dimensional model with a star schema structure to connect various hospital facility and ownership type data with quality scores and infection rates and payment information. The ETL pipeline performs efficient raw data extraction that proceeds to structural transformation before achieving data normalization for proper consistency. The system delivers practical analysis that reveals hospital performance indicators as well as ownership effects and infection patterns alongside geographic differences. How the visualizations display vital trends enables healthcare administrators to make better decisions through patient management.

The study revealed major differences in hospital quality ratings and financial outcomes and infection rates between hospital services and ownership organizations. The assessment highlights why healthcare policies together with resource distribution must depend on quantitative decision support for achieving enhanced clinical performance.

Future Improvements:

1. Enhancing Data Granularity: Incorporate patient-level data (de-identified) to analyze individual hospital stays, treatments, and outcomes for deeper insights.
2. Integrating Additional Data Sources: Include insurance claims, demographic data, and patient satisfaction surveys to provide a more comprehensive.
3. Predictive Analytics & Machine Learning: Apply ML models to predict hospital performance, patient outcomes, or financial risks based on historical trends.
4. Real-time Data Processing: Implement streaming ETL pipelines for real-time updates on hospital quality measures, ensuring timely decision-making.
5. Geospatial & Temporal Analysis: Introduce GIS-based visualizations to map hospital performance and analyze trends over time for better resource distribution.

By incorporating these improvements, the hospital data warehouse can evolve into a more robust, real-time, and predictive system, ultimately enhancing healthcare management, policy planning, and patient care outcomes.