Data Warehouse Implementation Report

1. Assignment Overview

The design and deployment of a data warehouse makes use of Python for storage purposes together with Snowflake for ETL tasks as well as analytics functions. Data acquisition processes along with transformations lead to modeling the data into a star schema structure which enables analytical querying and business intelligence analysis.

2. Data Sources

The assignment dataset originated from data.CMS.gov to present comprehensive Hospital Operations and Patients Data from various sources. The raw data analysis includes 100,000 records that present information about facility ID, facility name, measure name, state addresses and zip codes and other identical sections.

Dataset Overview

The hospital data warehouse uses different patient datasets to gather information about healthcare quality and performance measurement data. The United States-based facilities can access structured hospital data with quality metrics and payment information and infection rates through this database.

Potential Use Cases:

The The data source has multiple uses for business and healthcare intelligence purposes. These are potential use cases for the analysis:

1. Hospital Performance Benchmarking

   - The quality scores of hospitals can be evaluated against their national standards.
   - An analysis reveals the patient outcomes performance of hospitals at both highest and lowest ends.

2. Healthcare Cost Analysis

   - Data experts can study hospital payment practices through their ownership type classification as well as their geographical region and facility category specifics.
   - Senior executives should conduct financial efficiency assessments among various healthcare providers.

3. Identifying High-Risk Hospitals

- Hospitals demonstrating long-term low performance in complication rates as well as infection rates and mortality rates should be identified.
- The data serves to notify government bodies and insurance institutions about ways they can enhance healthcare standards.

4. Impact of Ownership and Hospital Type on Care Quality

- Researchers should compare the delivery of care services in hospitals which belong to for-profit entities, non-profit organizations and government facilities.
- The research investigates how academic medical centers and rural facilities and specialty care facilities impact the quality score measurement.

5. Predictive Healthcare Modeling

- The application of machine learning approaches helps forecast hospital performance through examination of past trends.
- Researchers must identify the essential elements which lead to either superior or inferior hospital rating scores.

6. Public Health Policy and Decision Making

- The information system delivers evidence-based data to healthcare policymakers who can use it for specific intervention strategies.
- Fund allocation guidance assists the funding of hospitals with subpar results to develop better care quality.

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.

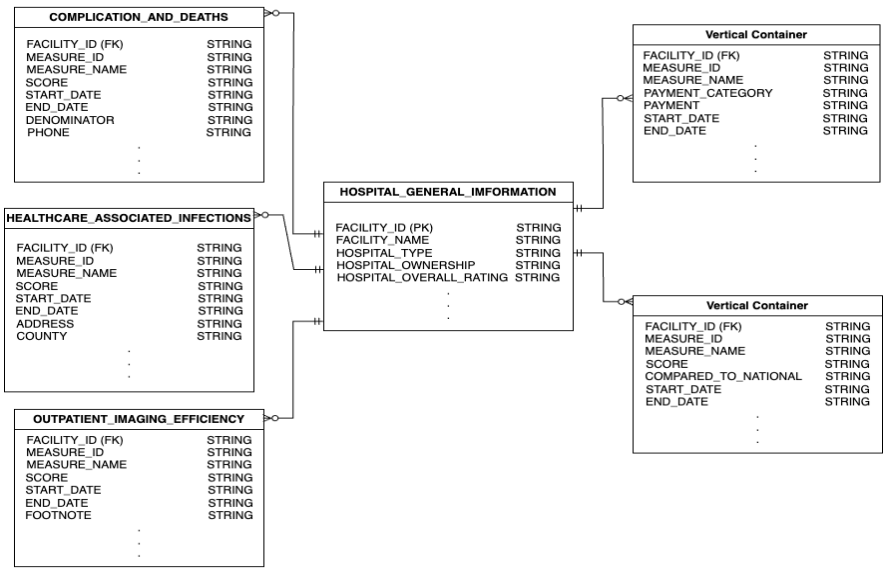
Healthcare Associated Infections	Healthcare_associated_infections.*csv*	Includes infection rates for hospitals.
Timely Effective Care	Timely_effective_care.*csv*	Measures timely care effectiveness.
Payment and Value of Care	payment_and_value_of_care.*csv*	Contains hospital payments and cost-related measures.

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

3. Normalised Database

Five specific tables exist in the normalized hospital data structure which connects to the central hospital_general_information table. The table architecture avoids data repetition by keeping facility characteristics such as name and location in the main HOSPITAL GENERAL INFORMATION table while individual performance metrics reside in separate COMPLICATIONS AND DEATHS, HEALTHCARE ASSOCIATED INFECTIONS, OUTPATIENT IMAGING EFFICIENCY, PAYMENT AND VALUE OF CARE and TIMELY AND EFFECTIVE CARE tables. The tables establish clean one-to-many relationships through their connection to hospital_general_information by using Facility_ID as a foreign key. The normalized methodology optimizes storage while minimizing anomalies during updates and provides adaptable querying between diverse measurement fields that upholds data reliability.

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:

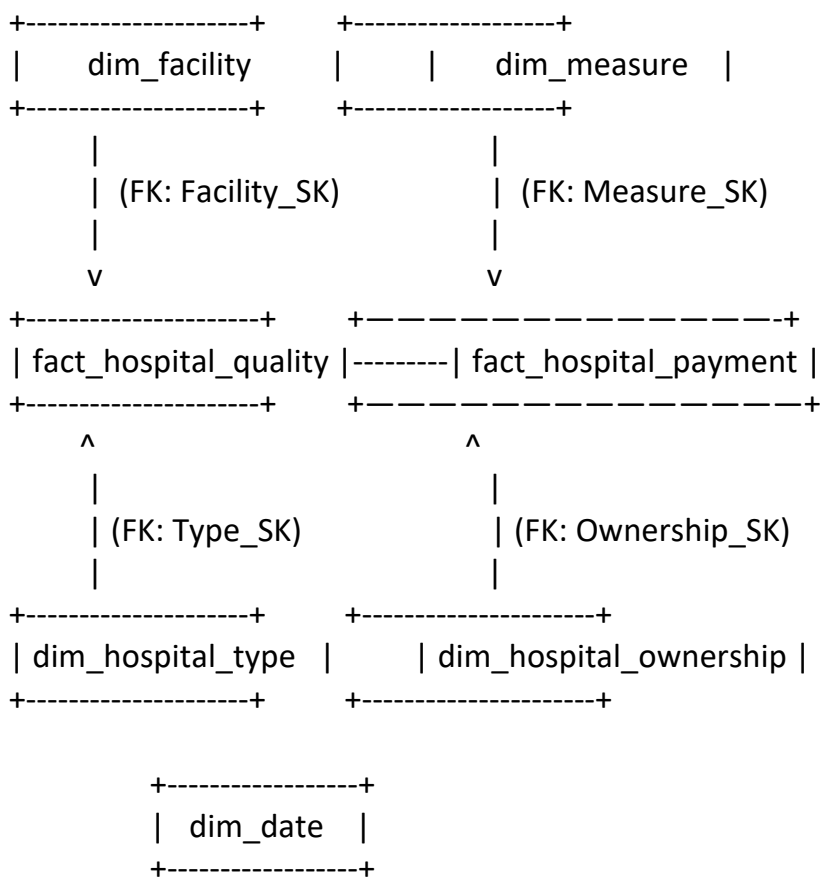
Table Name	Description
dim_facility	Contains hospital facility attributes like name, address, type, and ownership.
dim_measure	Stores the various healthcare quality measures being tracked.
dim_date	Holds date-related attributes for time analysis.
dim_hospital_type	Contains the different categories of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

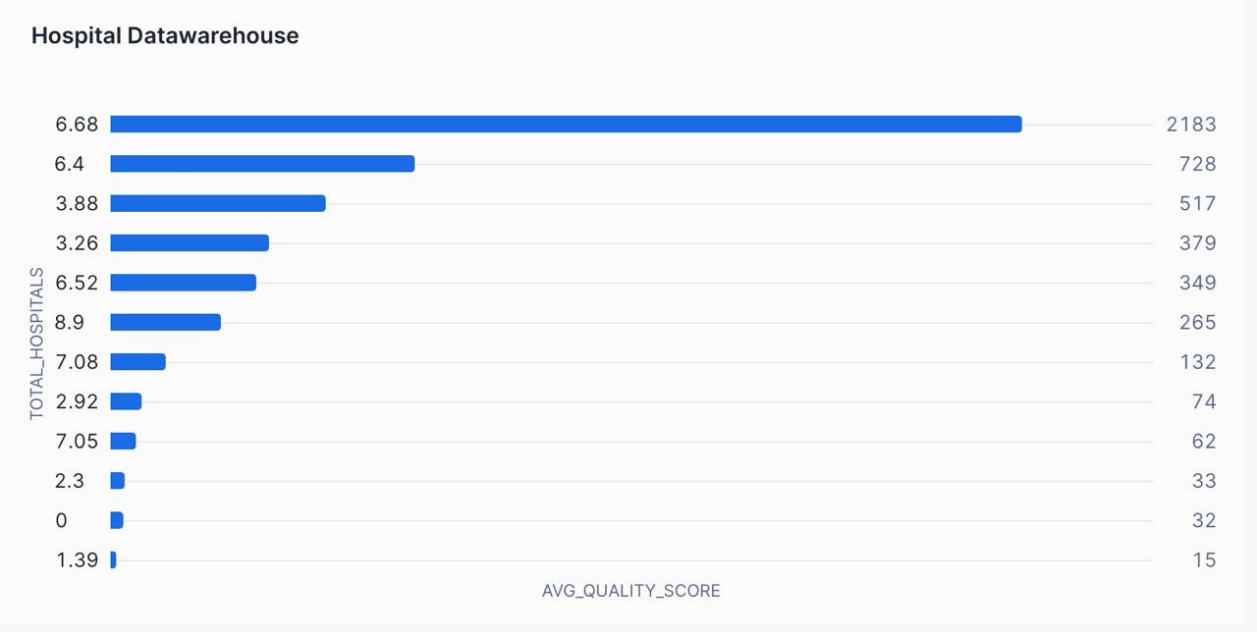
- dim_facility → To track which facility the payment data belongs to (Facility_SK).
- dim_measure → To identify the specific payment-related measure (Measure_SK).
- dim_date → To track when the payment was recorded (Date_SK).
- dim_hospital_ownership → To classify facilities based on their ownership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

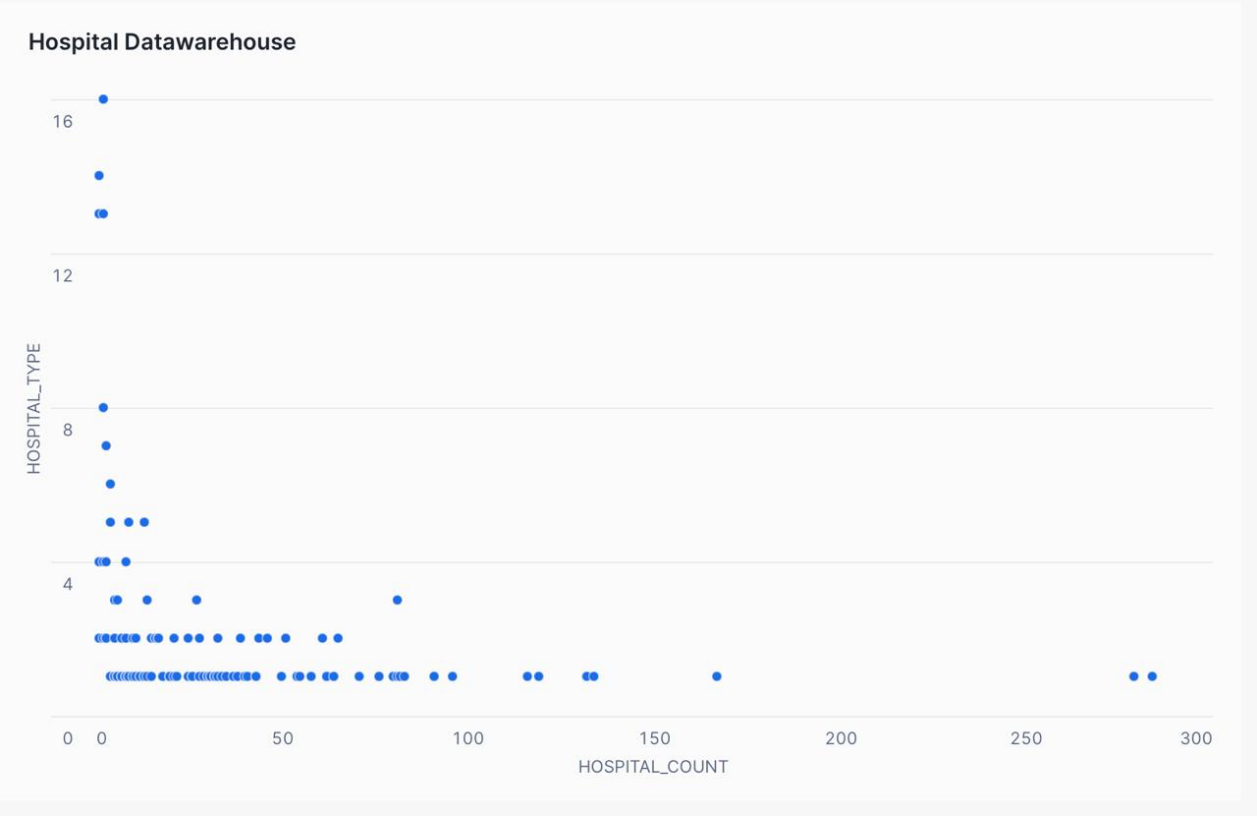
Business Insight: Identifies states with high infection rates, guiding policymakers and hospitals in improving healthcare safety and infection control.

```
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

An efficient integration of various health care datasets through the hospital data warehouse project enables supported analytical capabilities. We built an information model that utilizes star schema design to establish a connection between hospital facilities and ownership categories and their quality metrics and infection metrics and payment records. After extracting raw data using ETL pipeline structures transform the information to normalize it for consistent data storage. The system generates useful results that show performance metrics and both ownership influences and infection rates together with geographical differences. The visualization of essential trends are essential for healthcare administrators to create better decisions about patient management.

A study demonstrated substantial gaps appear regarding hospital quality ratings and financial outcomes and infection rates between hospital services and ownership organizations. Quantitative decision support stands essential for healthcare policy development and resource management to achieve superior clinical results according to the assessment.

Future Improvements:

1. More detailed patient information without personal identifiers should be included for analyzing specific patient hospital durations together with their received treatments and achieved outcomes.
2. Additional information sources should include insurance claims coverage data and demographic patterns at the same time as patient satisfaction feedback to enhance analysis capabilities.
3. Machine Learning models employed to forecast hospital operational measures and patient result evaluations and financial vulnerability with the aid of historical data analytics.
4. The organization should develop streaming ETL pipelines which provide real-time hospital quality measure updates for quick decision-making initiatives.
5. Geospatial & Temporal Analysis involves mapping hospital performance through GIS-based visuals to conduct time-sensitive resource distribution analyses.

The hospital data warehouse will strengthen its performance as a real-time predictive analytics system when these enhancements are implemented to deliver better healthcare management and policy planning and patient care outcomes.