🏥 Real-Time Healthcare ETL Pipeline – AWS

🧩**Project Goal:** Extract real-time healthcare data (e.g., vitals, device metrics), transform it to a standard format (cleaning, enrichment), and load it into Amazon Redshift for analysis and reporting.

1. Extract – Ingest Streaming Data

- **What Was Done:** Simulated a real-time data stream of patient vitals using a Python script. Data was sent to Amazon Kinesis Data Streams.
- **Tools:** Python, Boto3 SDK, Amazon Kinesis
- **Why:** To simulate continuous data from medical devices like heart monitors, glucose meters, or admission systems.

2. Transform – Clean and Standardize Data

- **What Was Done:** Used **AWS Lambda** to consume data from Kinesis and perform:
  - JSON parsing
  - Null handling
  - Unit normalization (e.g., Fahrenheit → Celsius)
  - Enrichment with patient metadata (e.g., age, gender) from DynamoDB
- **Tools:** AWS Lambda, Python, DynamoDB (for lookup), Pandas (optional for transformations)
- **Why:** Healthcare data often comes in inconsistent formats; transformations are essential before loading into analytics systems.

3. Load – Store in Amazon Redshift

- **What Was Done:**
  - Transformed data was batched and stored temporarily in **S3 (Parquet)**.
  - Used **AWS Glue job or Lambda** to load S3 data into **Amazon Redshift** using COPY command.
- **Tools:** S3, Redshift, Glue (or Lambda), SQL
- **Why:** Redshift provides fast, scalable querying for analytical dashboards and reports.

4. Schedule / Automate the Pipeline

- **What Was Done:**
  - Used **AWS Step Functions** to coordinate the flow:
    - Ingest stream → Transform → Stage to S3 → Load to Redshift
  - Enabled **retry logic** and **failure notifications**.
- **Tools:** AWS Step Functions, SNS for notifications
- **Why:** To make the ETL pipeline reliable, repeatable, and error-tolerant in production.

5. Visualize with Amazon QuickSight or Tableau

- **What Was Done:** Connected Amazon QuickSight to Redshift to build dashboards such as:
  - Patient admission volume by hour
  - Average heart rate by age group
  - Device failure trends
- **Tools:** Amazon QuickSight (or Tableau), Redshift
- **Why:** Business Intelligence teams and healthcare providers need visuals for quick decision-making.

Simulated Input Sample (Streamed into Kinesis)


```
{
 "patient_id": "P7654",
 "timestamp": "2025-06-12T15:45:30Z",
 "temperature_f": 101.2,
 "heart_rate": 130,
 "device_id": "D456"
}
```


Transformed Output (Loaded into Redshift)

patient_id	timestamp	temperature_c	heart_rate	age	gender
P7654	2025-06-12 15:45:30	38.4	130	65	Male

🎯 Outcome

- ETL pipeline processed and loaded healthcare data in near real-time (~minutes).
- Enabled analytics teams to run live queries on patient vitals and admissions.
- Reduced manual data wrangling and improved data accuracy.