xibtech
Powering Digital Growth

# Building a Secure Data Pipeline for a US Healthcare Analytics Company

Healthcare Case Study

xibtech
Powering Digital Growth

# Table of Content

xibtech
Powering Digital Growth

# Client Overview

A leading **US-based healthcare analytics company** providing population health insights to hospitals, insurance providers, and research organizations needed a **secure, scalable, and automated data pipeline**. The company processes millions of patient records daily from multiple EHR (Electronic Health Record) systems, wearable devices, and insurance claims data.

Because all this critical data was stored in different places and was not properly organised, their analysts couldn't get a clear picture. This made it very difficult to generate accurate reports for hospitals, track patient outcomes, or make data-driven decisions to improve operational efficiency.

xibtech
Powering Digital Growth

# Business Challenge

The biggest challenge was to create a single, secure, and unified source for all their healthcare data. They needed an automated system that could securely collect data from all sources, clean it, de-identify sensitive information, and prepare it for analysis. The goal was to build a modern reliable and automatic data pipeline.

- **Data silos and inconsistency**: Multiple healthcare systems stored data in different formats and locations.
- **Compliance**: Ensuring full HIPAA compliance for data storage and transmission.
- **Manual workflows**: ETL processes were semi-automated and prone to human errors.
- **Reporting delays:** Business intelligence dashboards took hours to refresh, limiting real-time decision-making.
- **Scalability**: Existing infrastructure couldn’t handle the growing volume of structured and semi-structured healthcare data.


xibtech
Powering Digital Growth

# Solution

We designed and built a complete, data engineering solution. Our approach was to create a clear, step-by-step process for the data to flow, get cleaned and anonymised, and become ready for reporting.

## 1. Collecting Raw Data (The Staging Layer):

- For files like CSV, Excel, and JSON containing patient and claims data, we set up a secure and simple process. The client's team would upload their files to a specific folder in AWS S3. An AWS Lambda function would automatically detect any new file, read it, and load the raw data into a secure staging area in our Snowflake data warehouse.
- For their provider activity data from Amplitude, we connected directly to its API. Our system would automatically pull the latest data every day and load it into the Snowflake staging area.

## 2. Cleaning and Transforming Data (The Silver & Golden Layers):

- Once all the raw data was in the Snowflake staging area, we used dbt (data build tool) to clean, transform, and model it.
- First, a set of dbt jobs would run to clean the raw data (like standardising medical codes, fixing dates) and move it to a 'Silver' layer. This data was clean but still very granular.

xibtech
Powering Digital Growth

# Solution

## 3. Creating Reports and Dashboards (The Presentation Layer):

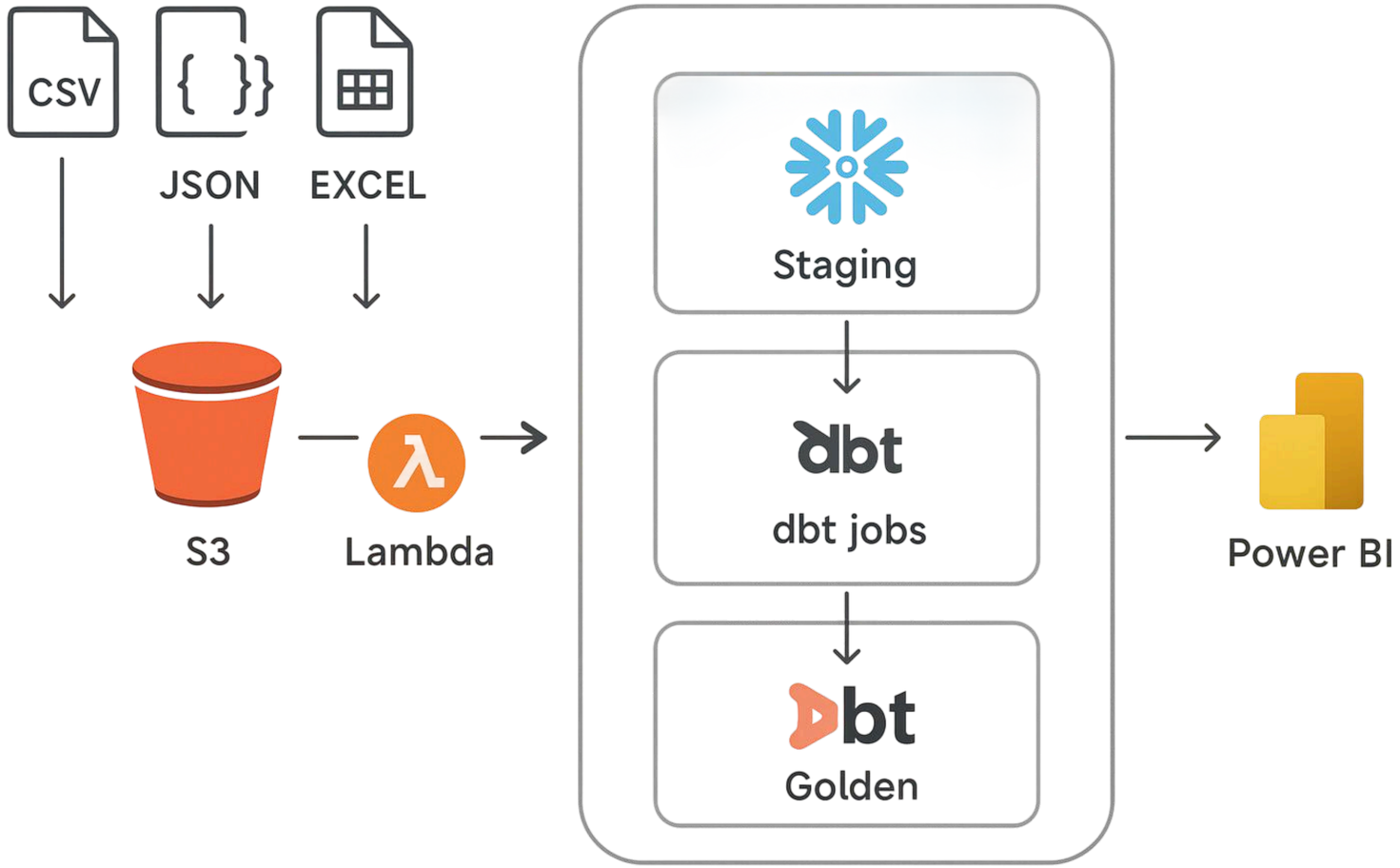
- The final, clean data in the Golden layer of Snowflake was now ready for analysis. We connected Power BI, a business intelligence tool, to this golden data.
- Using Power BI, we built easy-to-understand dashboards for their healthcare analysts. Now, they could easily track key metrics like patient admission rates, average length of stay, claim approval rates, and other important operational data in one place.

# Technology Stack

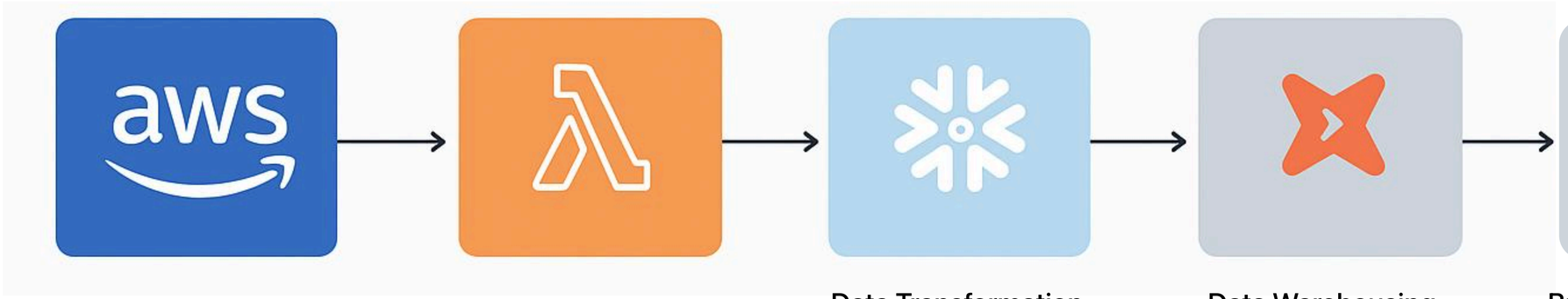
| Layer | Technology | Purpose |
| --- | --- | --- |
| Cloud & Storage | AWS S3 | Centralized, encrypted data lake for storing raw and processed healthcare data |
| Data Processing | AWS Lambda | Serverless compute for data ingestion, validation, and event-driven transformations |
| Data Warehouse | Snowflake | Scalable, secure warehouse for structured data analytics |
| Data Transformation | dbt (data build tool) | Modular SQL-based transformation and data modeling layer |
| Business Intelligence | Power BI | Interactive dashboards and embedded analytics for stakeholders |

xibtech
Powering Digital Growth

# Solution Architecture



# Solution Architecture




| Data Ingestion & Storage | Data Processing<br>AWS Lambda | Data Transformation<br>Snowflakes | Data Warehousing<br>DBT | Business Intelligence &<br>Analytics (Power BI) |
|---|---|---|---|---|
| • **Sources:** EHR systems, lab systems, wearable IoT data, and insurance claim APIs.<br>• **Ingestion:** AWS Lambda functions were triggered by events (e.g., new files in S3 or API calls).<br>• **Storage:** Raw data (CSV, JSON, HL7, FHIR) stored in AWS S3 buckets with encryption (AES-256). | • **AWS Lambda** validated schema conformity and anonymized patient identifiers using tokenization.<br>• Logging was integrated with AWS CloudWatch for traceability and monitoring.<br>• Lambda also triggered data movement from S3 raw to processed zones once validation was complete. | • dbt models transformed and joined data from multiple clinical and financial datasets.<br>• Transformation pipelines were modular and version-controlled via Git.<br>• dbt generated data lineage documentation for auditing and compliance.<br>• dbt Cloud handled CI/CD for transformation updates. | • Clean data loaded into Snowflake for structured querying.<br>• Role-based access controls and encryption at rest/in transit ensured HIPAA compliance.<br>• Time-travel and cloning features supported easy rollback and testing without impacting live analytics. | • **Power BI** connected directly to Snowflake for real-time querying.<br>• **Custom APIs** allowed embedding dashboards into the client's healthcare portal for stakeholders. |

xibtech
Powering Digital Growth

# Security & Compliance Implementation

- **Encryption**: Data encrypted both at rest (S3, Snowflake) and in transit (TLS 1.2).
- **Access Controls**: AWS IAM roles with least-privilege access policies.
- **Data Masking:** PII masked using Lambda scripts during ingestion.
- **Logging & Auditing:** AWS CloudTrail for full auditability of data access and changes.
- **HIPAA Compliance:** Ensured by aligning architecture with AWS HIPAA-eligible services.

# Performance Optimization

- Snowflake's **automatic scaling** and **query caching** improved report refresh time by 70%.
- Lambda-based event-driven ingestion reduced latency in data availability.
- dbt modular transformations shortened pipeline maintenance time by 40%.
- S3 lifecycle policies automatically archived old data to Glacier, optimizing cost.

# Outcomes & Business Impact

| Metric | Before | After |
| --- | --- | --- |
| Data Processing Time | 6 hours | 30 minutes |
| Dashboard Refresh | Once daily | Real-time (every 15 min) |
| Data Accuracy | ~80% | 99.8% validated |
| Storage Cost | High | 35% reduction via tiered storage |
| Compliance Risk | High | Fully HIPAA-compliant |

## Key Achievements

- Centralized data ecosystem for all healthcare operations.
- Real-time analytics empowered faster decision-making.
- Automated data quality checks improved trust in insights.
- Future-ready architecture capable of scaling with data growth.

# Future Enhancements

- Integrate **AWS Glue** for more complex ETL scenarios.
- Implement **ML models in Snowflake** for predictive analytics (e.g., patient risk scoring).
- Enable **data sharing** across partner healthcare organizations using Snowflake Data Sharing.
- Expand dashboard personalization in Power BI for role-based insights.

# Conclusion

The project successfully delivered a secure, **HIPAA-compliant, and fully automated data pipeline leveraging AWS S3, AWS Lambda, Snowflake, dbt, and Power BI.**

This implementation transformed the client's fragmented data landscape into a **modern cloud analytics platform**, empowering their analysts and executives to make data-driven healthcare decisions faster and with full confidence in data integrity.