Clinic Plus : A Data Engineering Capstone Project

By : Harshit Patel
Email Address : harshitpatel4ds@gmail.com
Linkedin Profile : Harshit Patel | LinkedIn
Github Account : Harry4ds

Table of Contents

Introduction

Healthcare providers today face an overwhelming challenge: managing vast amounts of data coming from a variety of sources while ensuring security, compliance, and accessibility. In Alberta's clinical environment, patient care relies heavily on accurate, timely, and well-organized information. Clinic Plus is an innovative cloud-based solution designed to address this

challenge. By leveraging Microsoft Azure's Lakehouse architecture, Clinic Plus centralizes data, enables real-time and batch processing, and delivers high-quality, analytics-ready datasets to a wide range of healthcare stakeholders.

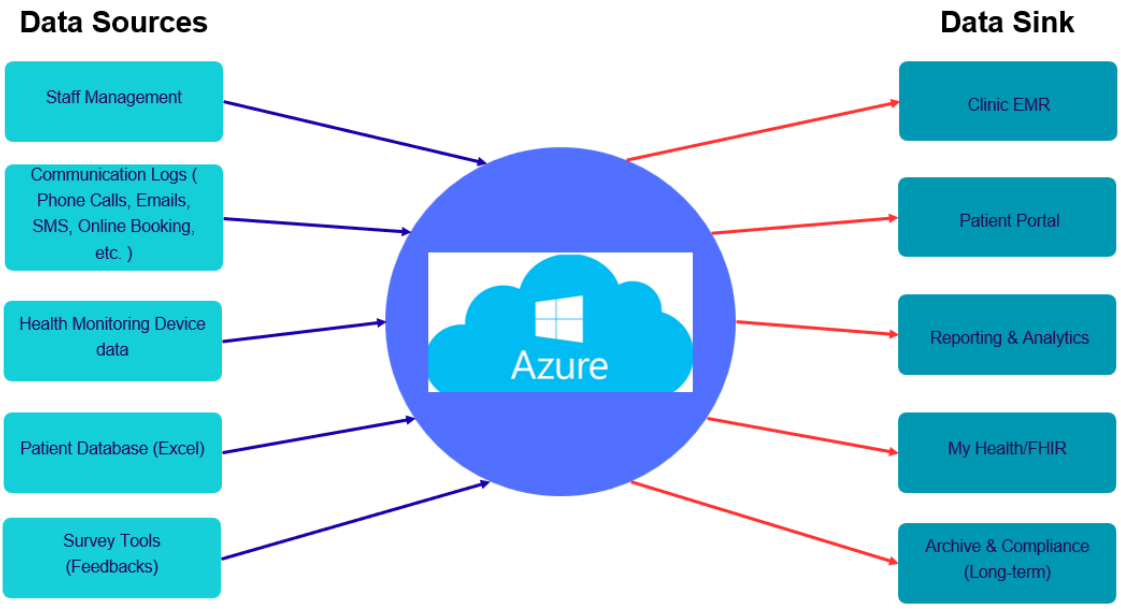
Objectives and Mission

The primary objective of Clinic Plus is to design and implement a scalable Azure-based data platform that unifies both clinical and operational data for medical clinics across Alberta. This unified platform ensures that decision-makers—from physicians to administrators—can access secure, up-to-date insights at any time.

Our mission is to create a data ecosystem that supports real-time insights, predictive analytics, and regulatory compliance. By implementing modern cloud practices, Clinic Plus aims to streamline workflows, improve patient experiences, and empower healthcare providers to make data-driven decisions.

Vision Diagram

Vision – Diagram of Clinic Plus



The diagram illustrates the high-level vision for **Clinic Plus**, showing how various clinical and operational data sources are centralized in an **Azure Cloud platform** and delivered to multiple end-user applications and systems.

Data Sources (Left Side)

1. **Staff Management** – HR and workforce-related data such as schedules, roles, certifications, and attendance.
2. **Communication Logs** – Patient interactions across channels (phone calls, emails, SMS, online booking platforms).
3. **Health Monitoring Device Data** – Continuous patient health metrics collected from connected medical devices.
4. **Patient Database (Excel)** – Structured patient records and demographic details stored in spreadsheet formats.
5. **Survey Tools (Feedback)** – Patient satisfaction and feedback data gathered through survey platforms.

Azure Cloud (Center)

All incoming data flows into the **Azure Cloud environment**, which serves as the central hub for ingestion, processing, storage, and security. This enables the Lakehouse architecture to unify both real-time and batch datasets.

Data Sinks (Right Side)

1. **Clinic EMR** – Electronic Medical Records for healthcare professionals to manage patient clinical data.
2. **Patient Portal** – Secure online access for patients to view appointments, test results, and personal health information.
3. **Reporting & Analytics** – Analytical dashboards and KPI tracking for operational and clinical insights.
4. **My Health/FHIR** – Standardized health data exchange using the FHIR protocol for interoperability with other systems.
5. **Archive & Compliance (Long-term)** – Secure long-term storage for regulatory compliance, audits, and historical data reference.

Overall Flow

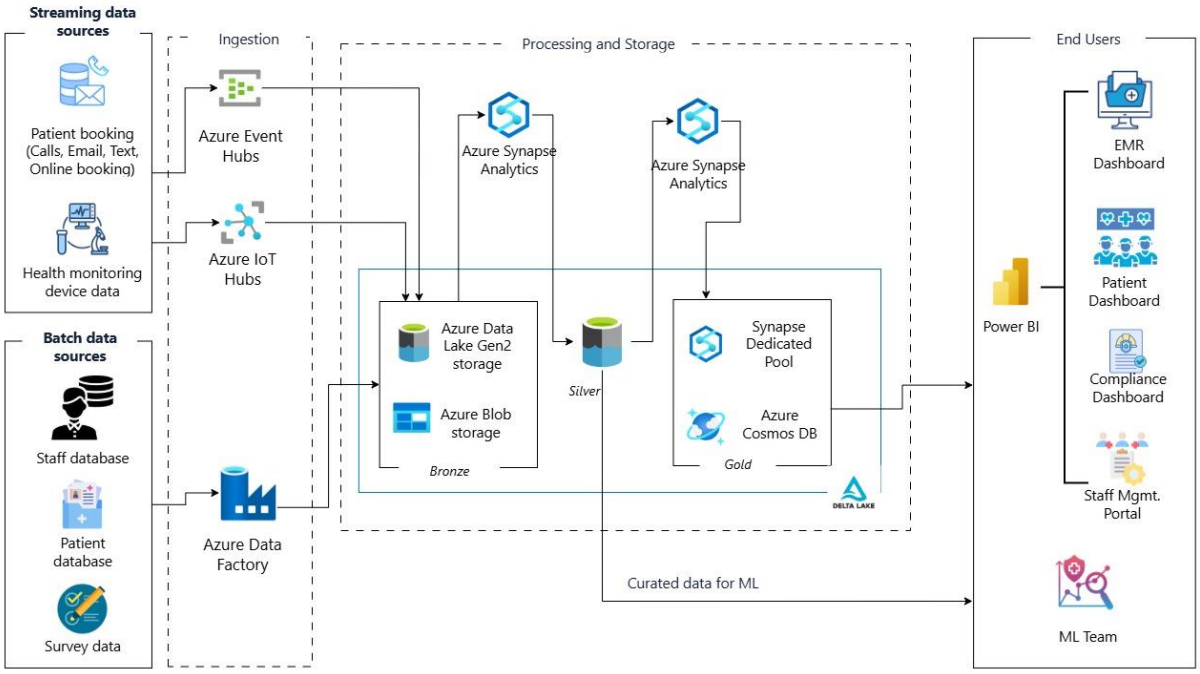
This vision emphasizes a **centralized Azure-based platform** that seamlessly connects diverse data inputs to relevant outputs, ensuring accessibility, interoperability, compliance, and data-driven decision-making across the clinic ecosystem.

Why Azure Lakehouse for Healthcare

The Azure Lakehouse architecture combines the flexibility of data lakes with the structured, query able nature of data warehouses. This hybrid model is ideal for healthcare environments, where raw data from devices and logs needs to coexist with curated datasets ready for reporting and analytics. The Lakehouse approach enables the ingestion of diverse formats—such as CSV, JSON, and Parquet—while maintaining scalability and performance.

In the context of Alberta's clinics, this means everything from patient monitoring device readings to staff schedules can be ingested, stored, processed, and analyzed in a unified platform.

Final Architecture



This architecture diagram illustrates the **end-to-end data flow** in the Clinic Plus Azure Lakehouse solution, from **data ingestion** to **end-user consumption**. It integrates **streaming data**, **batch data**, advanced **processing/storage layers**, and **role-specific dashboards**.

1. Data Sources

- **Streaming data sources**:
  - *Patient booking systems* (calls, emails, texts, online bookings)
  - *Health monitoring devices* that provide continuous readings such as vitals or diagnostics.
- **Batch data sources**:
  - *Staff database* containing personnel records, schedules, and certifications.
  - *Patient database* with clinical and administrative records.
  - *Survey data* collected periodically from patients or staff.

2. Data Ingestion

- **Azure Event Hubs** and **Azure IoT Hubs** handle **real-time ingestion** of continuous data streams from devices and booking systems.
- **Azure Data Factory (ADF)** manages **batch ingestion**, pulling large datasets from databases and survey tools on a scheduled basis.

3. Processing and Storage Layers

- **Bronze Layer** (*Azure Data Lake Gen2* and *Azure Blob Storage*):
  - Stores raw, unprocessed data in its original format (CSV, JSON, Parquet).
  - Data is immutable and partitioned for efficiency.
- **Silver Layer** (*Azure Synapse Analytics*):
  - Cleans, deduplicates, and normalizes data into curated tables for analysis and machine learning.
- **Gold Layer** (*Synapse Dedicated Pool* and *Azure Cosmos DB*):
  - Contains aggregated, analytics-ready datasets, including KPIs for dashboards.

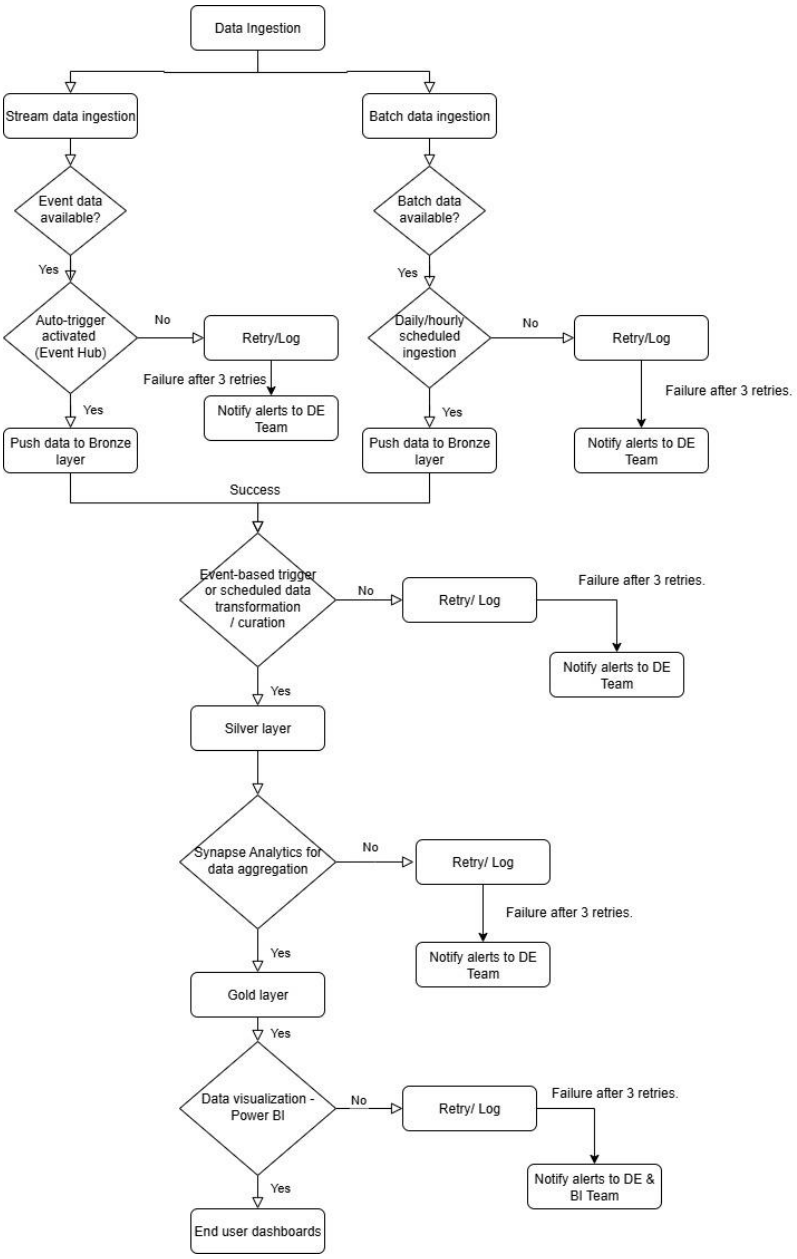
4. Data Consumption and Access

- **Power BI Dashboards**:
  - ***EMR Dashboard*:** For clinicians to review patient histories, vitals, and treatments.
  - ***Patient Dashboard***: For patients to view their health records and appointments.
  - ***Compliance Dashboard***: To monitor adherence to healthcare regulations.
  - *Staff Management Portal*: For HR and management to oversee staffing needs **and schedules**.
- **ML Team Access**:
  - Receives curated datasets for predictive modeling, anomaly detection, and AI-driven insights.

5. Machine Learning Integration

- The Silver Layer feeds curated datasets to the **ML Team**, enabling advanced analytics and predictive capabilities such as identifying patient risk factors or optimizing staffing.

Pipeline Execution Strategy



The data pipeline follows a structured, multi-layered execution flow designed to handle both real-time and batch data efficiently. Streaming data from patient booking systems and health monitoring devices is ingested through **Azure Event Hub** or **Azure IoT Hub** using auto-triggers, ensuring immediate capture of live events. Batch data from staff databases, patient records, and survey tools is ingested via **Azure Data Factory** on scheduled daily or hourly triggers.

All ingested data first lands in the **Bronze Layer** within **Azure Data Lake Gen2/Blob Storage** in its raw, immutable format. Event-based or scheduled transformations are then executed using **Azure Synapse Analytics Spark Pools** to clean, deduplicate, and normalize data into the **Silver Layer**, applying the parent-child structure for unified data models.

Aggregations and KPI calculations are performed in the **Gold Layer** using **Azure Synapse Dedicated SQL Pools** and, where applicable, **Azure Cosmos DB**, producing dashboard-ready datasets. The curated and aggregated data is visualized through **Power BI**, providing interactive EMR, patient, compliance, and staff management dashboards to end users, while also supplying curated datasets to the ML team.

At every stage of the pipeline, automated retry logic (up to three attempts) is implemented. If failures persist, detailed logs are generated and alerts are sent to the **Data Engineering** and, where relevant, **BI teams**. This ensures rapid issue resolution, minimal downtime, and reliable delivery of high-quality data to consumers.

Pipeline Fail Strategy

The pipeline integrates targeted failure-handling measures for each Azure component to ensure uninterrupted data flow and rapid recovery. **ADF Pipelines** employ a retry policy with alerts, monitored through Azure Monitor, ADF Alerts, and Activity Runs. For **Event Hub**, dead-lettered events are captured directly to Blob Storage, with failures tracked via Diagnostic Logs and Azure Log Analytics. **IoT Hub** leverages durable functions and queue backups, monitored through Azure Monitor, Application Insights, and IoT Hub Metrics.

During data processing, **Azure Synapse Analytics** (Spark pools and SQL pools) uses Try/Except blocks with logging to ADLS, timeout configurations, and automatic retries, monitored via Synapse Workspace Logs, the monitoring dashboard, and Spark Job Monitoring. Storage reliability is maintained in **Azure Data Lake Gen2** and **Blob Storage** by enabling Soft Delete and Versioning, with Azure Storage Metrics and Monitor Logs providing oversight. Finally, **Cosmos DB** employs throughput auto-scaling combined with a TTL strategy, monitored through Cosmos DB Insights and Diagnostic Logs.

This layered approach ensures that failures are detected early, mitigated automatically where possible, and escalated to the right teams with detailed diagnostics, maintaining consistent data availability across the bronze, silver, and gold layers.

Challenges

While the benefits of Clinic Plus are significant, the implementation journey comes with challenges:

- **Data Privacy & Compliance**: Ensuring compliance with healthcare regulations such as HIPAA and Alberta's Health Information Act.
- **Integration Complexity:** Unifying data from disparate systems, devices, and formats.
- **Real-Time Performance**: Maintaining low-latency data processing under high volumes.
- **Data Quality Assurance: Preventing** duplication, inconsistency, and incomplete records.

Conclusion

Clinic Plus represents a forward-thinking approach to healthcare data management in Alberta. By integrating Azure's Lakehouse architecture with a carefully designed ingestion and processing strategy, it delivers secure, timely, and insightful data to all stakeholders. The solution not only addresses current operational needs but also lays the foundation for future innovations, including AI-driven diagnostics, predictive staffing, and personalized patient care.