Data Architecture for Real-Time IoT Solutions in EMS Services Using Databricks and Apache Spark

Introduction:

This document outlines our proposed data architecture using Databricks and Apache Spark, designed to handle real-time data from IoT devices in ambulances. The architecture is divided into three layers—Bronze, Silver, and Gold—each crucial for transforming raw data into actionable insights. This setup supports the demanding data processing needs of emergency medical services.

Below is a visual representation of the architecture, illustrating the progression and processing of data across the Bronze, Silver, and Gold layers. This diagram helps clarify the structure and function of each layer, setting the context for the detailed discussions to follow.

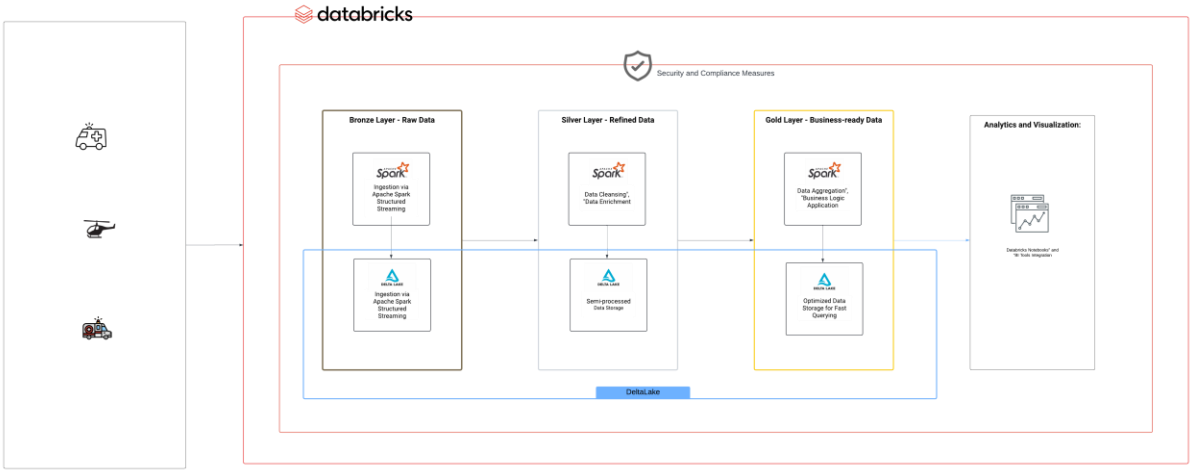

Figure 1: Overview of Data Architecture for Real-Time IoT Solutions

Bronze Layer: Raw Data Ingestion and Storage

Objective: The primary goal of the Bronze Layer is to capture and store raw data from ambulances without any processing, preserving its original form for traceability and potential reprocessing needs.

Processes and Tools:

- **Data Ingestion**: Utilises Apache Spark Structured Streaming to ingest real-time data from various sensors and OBD systems installed in ambulances. This approach handles high-frequency data efficiently, ensuring no data loss.
- **Storage**: Data is stored in Delta Lake, leveraging its robust capabilities to manage large volumes of raw data with high durability and reliability.

Data Types Managed:

- Sensor outputs including GPS, accelerometer readings, and engine diagnostics.
- Timestamps and device-specific identifiers for data traceability.

Silver Layer: Data Refinement and Integration

Objective: To refine and integrate the ingested data, making it more reliable and suitable for complex queries and analysis. This layer focuses on cleansing and aligning data from multiple sources.

Processes and Tools:

- **Data Cleansing**: Apache Spark is used to clean the data, removing inaccuracies and irrelevant information.
- **Data Enrichment**: Combines data streams, such as aligning GPS and accelerometer data based on timestamps, to enrich the data set.
- **Semi-Processed Data Storage**: Refined data is stored back in Delta Lake, optimised for better performance in data querying and retrieval.

Benefits:

- Ensures data reliability and consistency.
- Prepares data for advanced analytical processes and business applications.

Gold Layer: Business-Ready Data Transformation

Objective: To transform refined data into actionable business insights. This layer focuses on applying business logic and creating aggregated views suitable for decision-making and reporting.

Processes and Tools:

- **Data Transformation**: Utilising Apache Spark for advanced data transformations and aggregations, including calculating operational metrics and performance indicators.
- **Optimised Storage**: Storing transformed data in Delta Lake where it is structured and indexed for quick access, supporting real-time analytics and business intelligence applications.

Outcomes:

- Provides EMS providers and stakeholders with actionable insights for improved decision-making.
- Supports real-time operational adjustments and strategic planning.

Security and Compliance Overlay

Coverage: Spanning all layers, this overlay ensures that all data, regardless of its stage in the process, is handled with strict adherence to security standards and regulatory compliance.

- **Encryption**: Implemented at both data-at-rest and data-in-transit phases.

- **Access Control**: Role-based access controls are enforced to ensure that only authorised personnel have access to sensitive data.
- **Compliance**: Regular audits and compliance checks to align with industry standards such as GDPR.

The structured approach of the Bronze, Silver, and Gold layers within our data architecture ensures that data flows smoothly from collection to actionable insights. This architecture not only supports real-time data processing requirements but also adheres to stringent security and compliance standards, making it a robust solution for EMS services.

Documentation Strategy

The documentation should be comprehensive, well-organised, and accessible to all stakeholders. Here's I will approach it:

a. Technical Documentation

- **Architecture Diagrams**: Use tools like Lucidchart, Draw.io, or Microsoft Visio to create detailed diagrams illustrating the data flow, components, and interactions within the data architecture.
- **Data Dictionary**: Document each data element in the system, including its source, type, allowed values, owner, and any transformations it undergoes.
- **API Documentation**: If the architecture involves APIs, provide detailed documentation, including endpoints, request/response formats, and example calls, using tools like Swagger or Postman.
- **Configuration and Setup Guide**: Detail the setup process for the architecture, including hardware and software requirements, dependencies, and installation steps.
- **Operational Guide**: Include monitoring, alerting, backup, and recovery processes. Explain the steps to handle failures and data discrepancies.

b. Non-Technical Documentation

- **Executive Summary**: Create a high-level overview of the architecture's purpose, benefits, and how it aligns with business objectives. This should be understandable without technical expertise.
- **Use Case Documentation**: Describe how various roles within the company can interact with the system, the type of information accessible, and the benefits to each role.
- **Policy and Compliance Documentation**: Outline data governance policies, compliance with regulations (like GDPR), and how data security is maintained.

2. Communication Strategy

Effective communication ensures that all stakeholders understand the data architecture and its impact on the organization.

a. Initial Briefings

- **Executive Presentations**: Present the data architecture to C-level executives and board members focusing on strategic benefits, ROI, and alignment with business goals.
- **Departmental Meetings**: Tailor presentations for different departments (marketing, finance, operations), focusing on relevant features and benefits.