Data Pipeline Blueprint - Design Explanation

Beejan Technologies Customer Complaint Management System

What I Have Built

I have designed a comprehensive end-to-end data pipeline architecture to solve Beejan Technologies' customer complaint management challenges. The pipeline integrates four distinct data sources (Twitter, SMS, phone logs, and web forms), processes them through a unified transformation layer, stores data in a dual-layer architecture, and serves insights through multiple channels.

Architecture Overview

The pipeline follows a modern data engineering pattern with five core components:

1. Data Sources Layer My design captures complaints from four different channels that represent the complete customer feedback ecosystem:

- **Twitter/X**: Real-time social media complaints requiring immediate attention
- **SMS Messages**: Text-based complaints processed in hourly batches
- **Phone Call Logs**: Structured call center data processed daily
- **Web Forms**: Online complaint submissions captured in real-time

2. Ingestion Layer I implemented a hybrid ingestion strategy that handles both real-time streaming and batch processing. This approach recognizes that different data sources have different velocity requirements - social media complaints need immediate processing for reputation management, while phone logs can be processed in daily batches for operational efficiency.

3. Processing Layer The processing component includes three critical stages:

- **Data Cleaning**: Standardizes formats across sources, removes duplicates, and handles missing data
- **Metadata Validation**: Ensures data quality through schema validation and business rule enforcement
- **Data Enhancement**: Adds business value through complaint categorization, sentiment analysis, and customer data enrichment

4. Storage Architecture I chose a dual-layer storage strategy that balances flexibility with performance:

- **Data Lake (JSON format)**: Preserves raw data for reprocessing, compliance, and backup purposes
- **Data Warehouse (Parquet format)**: Optimizes processed data for analytics and reporting

5. Serving Layer The final component delivers data through four channels:

- **Analytics Dashboards**: Real-time monitoring and trend analysis
- **API Endpoints**: System integration and application feeds
- **Machine Learning Models**: Predictive analytics and automated classification
- **Report Generation**: Automated business reporting

6. Orchestration Layer I added orchestration to manage the entire pipeline workflow, including scheduling, error handling, monitoring, and performance optimization.

This blueprint provides Beejan Technologies with a robust foundation for transforming their customer complaint management from a manual, siloed process into an automated, integrated system that delivers real-time insights and supports data-driven decision making. The design choices reflect both current business needs and future growth requirements, ensuring long-term value from the implementation investment.