**📘 Data Pipeline Design for Beejan Technologies**

**📖 Project Overview**

Beejan Technologies receives thousands of daily customer complaints through different channels such as social media, call centers, SMS, and website forms. Currently, the data is stored in multiple formats, reports are manually compiled, and teams work in silos, leading to delays and inefficiencies.

This project provides a **conceptual design of an end-to-end data pipeline**. The pipeline is focused on unifying, cleaning, enriching, and organizing customer complaint data to make it ready for analysis, reporting, 43and decision-making.

**🛠️ Conceptual Pipeline Design**

The pipeline has **seven main steps**:

1. **Source Identification**
   * Complaint data comes from multiple channels: social media (streaming), call center logs (batch), SMS (batch), and website forms (batch).
2. **Ingestion Strategy**
   * Data is ingested through APIs, log files, and streaming feeds depending on the channel.
   * Real-time data (e.g., Twitter) is captured continuously.
3. **Processing / Transformation**
   * Raw data is cleaned, standardized, and enriched.
   * Complaints are categorized (e.g., network issues, billing errors, customer service).
4. **Storage Options**
   * A **Data Lake** stores raw unprocessed data.
   * A **Data Warehouse** stores cleaned and structured data for reporting and analysis.
5. **Serving**
   * Clean data is made available through:
     + **Dashboards & KPIs** for managers.
     + **Analyst Queries** for deeper insights.
     + **Alerts & Notifications** for urgent or abnormal issues.
6. **Orchestration & Monitoring**
   * The pipeline runs on a scheduled and event-driven basis.
   * Failures are detected and alerts are sent automatically.
7. **DataOps**
   * The pipeline is designed to run in a production environment.
   * It is scalable to handle large complaint volumes daily.

**✅ Assumptions**

* Social media complaints (e.g., Twitter, Facebook) are treated as **real-time streaming data**.
* Call center logs, SMS records, and website forms are ingested as **batch data**.
* The company already has the infrastructure to store both raw and cleaned data.
* Complaint categories are predefined (network, billing, customer service, etc.).
* Alerts are required only for urgent complaints (e.g., outages, repeated billing errors).
* The reporting team uses dashboards, while analysts need direct query access.

**🎯 Challenges / Unknowns**

* The exact **volume of data** is not specified (may affect scalability design).
* The **accuracy of complaint classification** may depend on the quality of data.
* Real-time social media monitoring may need advanced filtering to avoid noise.

**📊 Expected Outcomes**

* A **unified data pipeline** that connects all sources into one system.
* Faster, automated reporting instead of manual spreadsheets.
* Actionable insights through dashboards, alerts, and analyst queries.
* Improved decision-making and quicker response to customer issues.