BEEJAN TECHNOLOGIES CONCEPTUAL END-TO-END DATA PIPELINE EXPLANATION.

1. Design Choice

This pipeline is designed to integrate multiple customer complaint channels like social media, call center log files, SMS, and website forms into a unified system. Since these data sources differ in format and frequency, a hybrid ingestion approach is chosen: streaming for real-time sources like social media and batch/micro-batch processing for periodic uploads like call center log files, web forms & SMS.

After ingestion using API feed file upload and streaming events, the data passes through processing & transformation stage where cleaning and standardization **(**formats are unified, and timestamps normalized), deduplication (duplicates are removed) categorization (complaints are grouped into different categories; poor network, incorrect billing, or bad customer service), & PII masking (obscuring Personally Identifiable Information to enhance data security).

After cleaning, data is stored in a data lake and a data warehouse. Data lake for unstructured/semi-structured data (JSON) and a data warehouse for structured data (CSV). This separation allows both historical archiving and optimized querying. These data are stored in efficient, query-friendly formats to balance performance and storage efficiency.

The serving layer supports both analytical queries by business/data analysts and dashboards for managers. This ensures both ad-hoc exploration and standardized reporting are supported.

Finally, orchestration and monitoring ensure that pipelines run at the right frequency, with scheduling (automated checks) for data delays, ingestion failures, and data quality issues. A DataOps framework provides CI/CD for pipeline updates, monitoring, and governance to ensure production readiness.

2. Assumptions / Thought Process

i. Social media complaints arrive continuously; therefore, streaming ingestion is appropriate.
ii. Call center log files and website form submissions are stored in files or structured databases, which can be processed in scheduled batches.
iii. SMS messages arrive in near-real time but may be buffered in micro-batches for efficiency.
iv. Business users need both real-time visibility and historical analysis which drives the choice of both a data lake and warehouse.

3. Challenges / Unknowns

i. Streaming data and batch data must be reconciled into consistent reporting periods.
ii. Sensitive customer data in logs and SMS must be handled with privacy considerations, requiring masking.

iii. Volumes of social media data may spike unpredictably, and the pipeline must adapt without delays.

4. Additional Notes

i. The end-to-end pipeline ensures complaints are captured, standardized, enriched, stored, and served across teams.

ii. The presence of a DataOps layer emphasizes reliability, repeatability, and production readiness, ensuring the pipeline can evolve as business needs grow.