**Conceptual End-to-End Data Pipeline for Beejan Technologies**

The purpose of the pipeline is to systematically convert disorganized customer grievances into actionable intelligence for the decision-makers. The objective is to capture every possible data interaction within the company to streamline data for the entire firm.

**Data Sources:** I focused on four main types of data—social media, call center logs, SMS complaints, and web forms. It is a mix of structured and unstructured data, and that is the reason the design incorporates extensibility.

**Ingestion:** Streaming and batch ingestion both apply in this scenario. Call center logs are best handled in batches (daily/hourly), while social media and SMS are more time-sensitive and require real-time streaming. The same is true for website forms, but we can view them as micro-batches or nearly real-time.

**Processing/Transformation:** To ensure data functionality, it necessitates removing duplicates, fixing format discrepancies, aligning timestamps and encodings, and sorting complaints into categories like network, billing, or customer service.

**Storage:** I opted for a dual-layer strategy: a data lake for holding all information in its original state and a data warehouse for the processed and organized data. This allows analysts to refer back to the original data if necessary, while reporting teams receive quick, prepared datasets. Columnar storage formats are optimal for efficiency.

**Serving:** The processed data will be accessible via a query layer that facilitates dashboards, automated reports, and on-demand queries. This assists various teams—from executives needing summaries to analysts seeking trends.

**Orchestration & Monitoring:** Scheduled batch jobs will execute at specific times, whereas streaming jobs operate continuously. Failures will be identified via monitoring systems that analyze logs, volumes, and rates of errors. Notifications will guarantee the team reacts swiftly.

**DataOps:** To maintain the reliability of the pipeline, it will operate in a scalable production setting with a distinct division between development, testing, and operational systems. Modifications will be managed with version control, verified through testing, and deployed via automated systems.

**Assumptions & Thought Process**

While designing this pipeline, I made a few assumptions:

1. Social media platforms and SMS providers allow access via APIs.
2. All sources can be aligned into a common schema (with fields like complaint type, customer ID, and timestamp).
3. Management wants both **real-time insights** (e.g., sudden spikes in complaints) and **historical analysis** (e.g., complaint trends by region over a year).
4. The pipeline should prioritize simplicity at first but remain flexible to scale in the future.

My thought process was to start broad by collect everything in raw form and gradually refine by cleaning and enriching the data until it is valuable enough for decision-making.

**Challenges & Unknowns**

There are a few open questions and risks that would need clarification before building this in real life:

* **Data integrity:** social media and SMS are chaotic. Identifying sarcasm, slang, or abbreviations may impact the accuracy of classification.
* **APIs and access:** There may be limitations on the use of social media data, and integrating SMS might incur additional charges or constraints.
* **Categorization of complaints:** Establishing precise categories will necessitate input from business stakeholders. Lack of alignment can result in classifications that do not meet managers' expectations.
* **Latency needs:** It is unclear if the management requires actual real-time dashboards (seconds) or if “near real-time” (minutes) is sufficient. This may alter design priorities.

In conclusion, the focus is on creating a process that integrates isolated data, refines it, and presents it in an organized, usable format. The integration of a data lake and a warehouse provides both adaptability and efficiency. Incorporating monitoring and DataOps practices emphasizes that this pipeline is practical and intended to be sustainable in a production environment.

This pipeline offers Beejan Technologies a unified source of truth for customer complaints, converting frustration into valuable insights. With time, it may develop further, such as by integrating sophisticated machine learning for sentiment analysis or predictive analytics for customer retention.