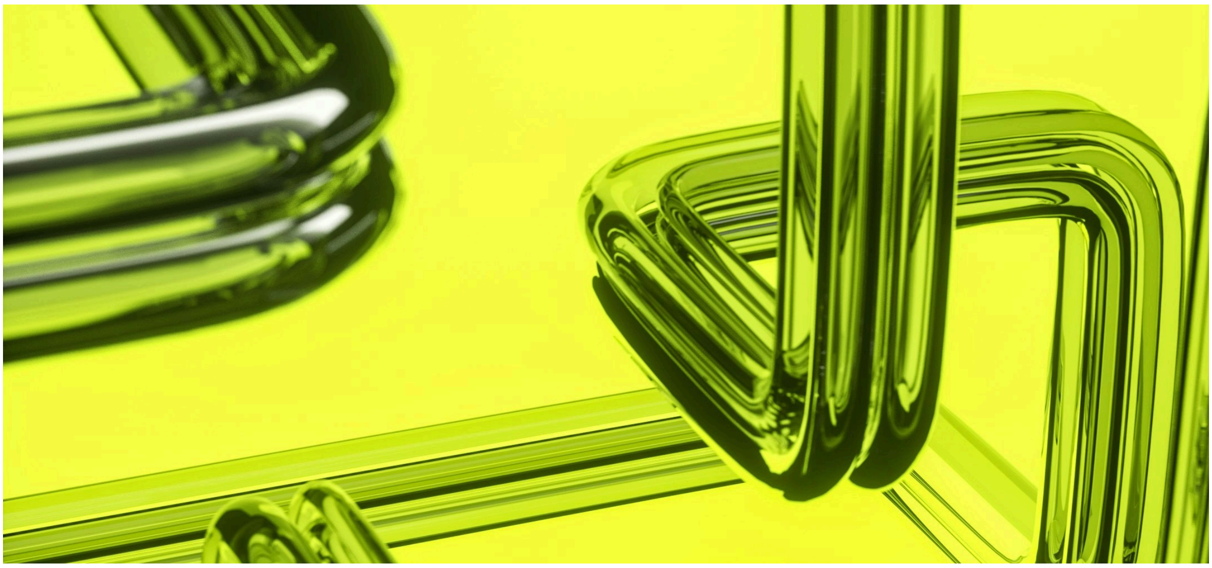

DE Fundamentals Assignment

Data Pipeline Design

```
https://lucid.app/lucidspark/a643a655-cf9b-4bda-acbd-72a540a0a7eb/edit?viewport_loc=-6332%2C-1123%2C22239%2C10682%2C0_0&invitationId=inv_48ed13af-0db7-4c51-aee2-0d9a0ea81b72
```

Written Explanation

Design Choices

- **Sources:** Multiple sources with varying formats (structured, semi-structured, and unstructured). Assumption: Call center logs are batch files, social media is streaming, and SMS/website forms are near real-time.
- **Ingestion:** Both batch and streaming ingestion to handle frequency differences.
- **Processing/Transformation:** Standardize formats, remove duplicates, and enrich with metadata. Complaints are classified using keyword-based tagging or rules.

- **Storage:** A layered storage approach, with raw data stored first, then processed, and finally curated for analytics. Both data lake and data warehouse are assumed.
- **Serving:** Data served to reporting dashboards, query interfaces, and potentially machine learning models for predictive insights.
- **Orchestration & Monitoring:** Pipeline scheduled (daily + streaming where needed). Alerts are set up for failures or delays.
- **DataOps:** Pipeline runs in a controlled environment, deployed in production with versioning and access control.

Assumptions

- Social media complaints are received via APIs, logs are captured in batch files, and SMS/website forms are stored in structured feeds.
- The company wants near-real-time insights for urgent issues (network outages) but daily aggregation for reporting.
- Teams will align on standard complaint categories.

Challenges/Unknowns

- Data quality may vary significantly (misspellings, incomplete complaints).
- Volumes from social media could spike unexpectedly.
- Alignment across business teams on categories and metrics may take time.
- Latency expectations (how “real-time” does management want?) need clarity.

Other Information

- The design focuses on concepts, not specific tools.
- This blueprint ensures scalability, flexibility, and better collaboration across teams.