BEEJAN TECHNOLOGIES CONCEPTUAL END-TO-END DATA PIPELINE REPORT

SUMMARY

Beejan Technology is a telecommunication company currently looking for a solution into building a data pipeline for handling and storing their customer's complaint that comes from different sources like social media, call center log files, SMS, and website forms which raises a huge concern. The goal of my report is to design a conceptual end-to-end solution that solves Beejan's problem from building a scalable, efficient and reliable pipeline to serving it to the downstream users such as the data analyst and the data scientist.

DESIGN CHOICE AND THOUGHT PROCESS

1. Source Identification and Ingestion Strategy: At this level, data is coming from different sources such as social media, call center log files, SMS, and website forms. I have decided that social media posts and website forms will be ingested via API (because they are generated online), call log files via csv (because I can easily store call log as csv) and SMS will be ingested from the database (because our sms logs are stored as a .db file). These data will be ingested using both batch (for data that can be scheduled) and streaming (to handle real-time request like payment and verification) processing.
2. Processing/Transformation: For the processing and transformation, these data will be standardized considering different aspect which include; creating a uniform data (data that is consistent and also comply with a specified format and structure e.g. making all date related data adhere to the date format), making data consistent by handling data validation (what is coming from the source should match the output) and also eliminate ambiguity to prevent duplications and errors when the data is being used. Most importantly, I will take time to understand the types of data. This will help give a better understanding on how to transform and process them. For the categorization of complaints, a sentimental analysis can be carried out to identify keywords in separating data.
3. Storage Options/ Serving: Every data, when ingested, will be stored in one central location at first (Data Lake). This is to ensure that I capture every single data from the source without leaving any behind even though I might not use all at the time but will still be stored for future purpose when needed. After transformation and we have a cleaned data, the needed data will be moved to a data warehouse for the purpose of serving to the downstream users. Although, data will be available in different format for different users (for example, non-technical people like stakeholders and the customers will prefer data to be presented with visuals like charts and trends, semi-technical people like the finance team will use spreadsheets for reconciliation, budgeting and preparing income statement while technical people from departments like Innovation and the data analyst will require data in a data warehouse or an API they can integrate to their applications or dashboard). Data scientist might get served in

parquet format because they are good for model creation and thorough analysis. Data will also be loaded incrementally to not repeat what has been stored already for efficiency and optimization.

4. Orchestration and Monitoring: Orchestration is very key to keep my pipeline running and active without my intervention. For batch processing of data like complaints from customer that are not urgent (e.g. total transaction of the day, total website visits), they can be scheduled to run at the end of business daily. For complains that needs quicks attention, a notification will be set to trigger when these occurs for quick attention (e.g. transaction failure, server shutdown, website down and other network issues).
5. DataOps: The pipeline will run both on the cloud and on premise for redundancy sake. If there is a data loss or an infrastructural issue within the organization, we can rely on the cloud for backups and to manage cost, using both methods will be the best option.

CHALLENGES (KNOWN AND UNKNOWS)

Known Challenges:

1. Data quality issues like duplicates and missing data is a top issue. Identifying and tackling this issues might require contacting some customers
2. Since data is coming from different sources, integration into a central storage is a big concern.
3. Categorizing data gotten from some sources like twitter is a challenge because they are not well structured.
4. Storage issue might occur if too many complaints are being reported and the database can’t accommodate them all

Unknown Challenges:

1. There might be a rise in the volume of complaints that we are not aware of yet.
2. Introduction of a new data type and getting complaints from a new source (not earlier introduced e.g. physical complain or contacting a staff personally to lodge complaint) might break the pipeline and pose threat to data completion that affects decision making.
3. Change in company’s policy or federal government policy might arise and affect the structure.