Objectives of the Project

Beejan Technologies is faced with lack of a pipeline to accommodate flow of data in the organization. Data from customers' complaints are stored in different formats and make analysis and reporting difficult. To overcome this challenge, the project aimed at providing a conceptual end-to-end data pipeline that will bring all these data together, transform the data, it ready to provide actionable insights.

Design Choices

Our choice of design for the pipeline will be **batch processing**. This choice is influenced by the nature of the data sources being large volumes of data coming in daily and do not require immediate analysis. The complaints from the customer will be scheduled to be ingested from the various sources in interval. In case of real-time data from social media (e.g. Twitter), the data can be collected on a scheduled basis using Twitter API.

Sources

The data come in different formats which include social media API, log files, SMS, and website forms. Our conceptual pipeline will be designed such that it will support a variety of data sources, including both structured and unstructured data.

Data Ingestion

This step involves the process of collecting the data from the different sources and moving it to our target destination. The ingestion of data from these varieties of sources will be done in batches (say every 6 hours, 12 hours, or 24 hours). The ingestion of data will involve:

- File uploads: Complaints that was channeled from call centers are stored in files. Also, data collected via website forms are either stored as csv or text files.
- API ingestion: Social media complaints (for example on Twitter) will be ingested via API. The real-time data from twitter can be collected in batches on a schedule interval via Twitter API. For example, we can fetch a n number (say 100) of tweets for every 1-hour timeframe. Data from SMS can also be ingested via API or third-party platform,

Data Transformation

The ingested data will then be transformed by performing some data cleaning, standardization, and aggregation. Unwanted information from log files, social media texts and so on will be removed. Transformation will also involve breaking down complaints into categories by normalization. Unstructured data will be converted to structured format that can then be load into data warehouse.

Data Storage

The raw data coming from the different sources can be stored in a data lake. However, the transformed data that has been structured will be stored in a data warehouse.

Data Serving

The transformed data already loaded in the data warehouse will be made accessible for the downstream users. The data is already in a suitable format that can be queried and serve the purpose of analytics, machine learning, reporting, decision-making and other business use.

Data Orchestration and Monitoring

The pipeline designed will be automated to run on scheduled intervals and include monitoring alerts in case of failures. The monitoring can include logging each stage of the pipeline in order to detect a stage that is breaking in the pipeline. Error messages should be logged into a file for easy debugging of codes.

DataOps

To make the pipeline available in production, the choice of platform is an important consideration. Appropriate testing will be done before moving the pipeline into production.

Challenges

Some of the challenges likely to be faced with project include:

- Handling real-time data streaming from social media.
- Little knowledge about running pipeline in production.