Group Name: Data Glacier Project Group

Name:	Ameya Damle	Divya Medapureddy
Email:	ameyadamleuk@gmail.com	
Country:	United Kingdom	USA
College:	University of Reading	University of Maryland Baltimore County
Specialization:	Data Science	

Problem Description

XYZ company is collecting the data customer using google forms/survey monkey and they have floated n number of forms on the web. Company wants to create a pipeline which will collect all the data of these google forms/survey monkey and visualize the data in the dashboard.

Company wants clean data and if there is any data issue present in the data then it should be treated by this pipeline (duplicate data or junk data). dedup check should be performed on the email id of the customer.

Business Understanding

Data Collection:

- The company is using Google Forms and SurveyMonkey to gather customer data.
- Multiple forms are being deployed on the web, generating various sources of data.
- Understanding the structure and content of these forms is crucial for effective data extraction.

Data Quality:

- The company requires clean and accurate data for analysis.
- Duplicates and junk data are potential issues that can affect the quality of analysis.

Pipeline Development:

- A data pipeline needs to be designed to automate the process of collecting, cleaning, and preparing the data for analysis.

Deduplication Strategy:

- Duplicate data is a concern, particularly for customer information.
- The pipeline should implement a deduplication strategy using email IDs as a unique identifier.

Data Visualization:

- The goal is to visualize the collected and cleaned data through a dashboard.
- Visualization enhances data interpretation and supports decision-making.

Scheduled Data Export:

- To set up a batch job using scheduling tools like cron jobs or dedicated ETL platforms like Apache Airflow.
- To schedule the batch job to run at specific times to automatically export the cleaned data into a master file or data lake.

Documentation:

- Maintain a documentation log where you record challenges encountered during the implementation.
- Detail the nature of each challenge, the steps taken to address it, and the final resolution.
- Include insights and lessons learned from overcoming these challenges.

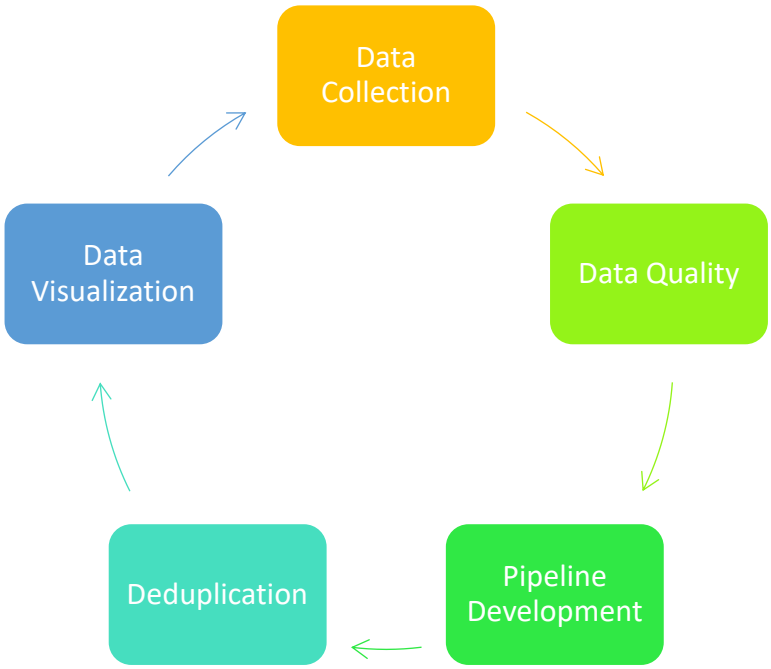

Figure 1: Data flowchart