DATAPIPELINE FOR DATA ENGINEERING

Data Sources

ETL Transformation

Data Collection and Ingestion

Data Preprocessing

Datawarehouse

(Big query Table)

Raw Data Storage

Processed Data (Staging Layer)

Fetching the data for analyzing

Visualization

Data Ingestion and Collection:

1. Mechanism should be built to handle heterogenous sources
2. There should be alert mechanism once file is received, or the files are delayed from the vendor.
3. There should be alert mechanism to handle different types of file format and handle.
4. There should be alert mechanism if the file or the data is not available in desired location in specific file location.
5. In case of aggregated data loading, there should be mechanism to check for new data and process the data once the new data is available.

DATA PREPROCESSING:

The raw file should be preprocessed to be consumed by the ETL.

1. The file should be placed in a specified directory to be consumed by the ETL directory.
2. The file should be archived for future reference.
3. The following sanity checks should be performed so that the file does not fail during the ETL process.

IBC checks for the files:

* Check for headers in the files
* Check the checksum of the file
* Check if we have received the full file or not
* (More checks could be included as per the format and type of the file decided by the business or the external vendor)

ETL PREPROCESSING

1. The raw file will be transformed and loaded into the respective staging table or the main table.
2. Proper dependencies need to be set up between the jobs to establish a relationship between the tables.
3. A separate pipeline needs to be created to handle the reject records
4. An alert system needs to be created for scenarios where the job fails alerting the impacted systems.
5. A alert mechanism needs to created incase the jobs doesn’t complete on time. The alert mechanism could be set up according to the SLA.