ETL process

1. Setting up Google Drive and Google Cloud Storage (GCS):
    - Activate the Drive API in Google Cloud Platform (GCP) to access data from Google Drive.
    - Create credentials to authenticate the API requests.
    - Create a service account and generate a key for it.
    - Make sure you have a bucket created in GCS to store the data.
2. Sending data from Google Drive to Google Cloud Storage:
    - Once you have the necessary credentials and bucket, you can use code to transfer data from Google Drive to GCS.
    - This step involves utilizing the activated Drive API and authenticating the requests using the service account credentials.
    - Ensure that the data from Google Drive is successfully transferred and stored in GCS.
3. Transforming data using Dataflow Workbench:
    - Utilize the Dataflow Workbench, which is a notebook environment, for data transformation tasks.
    - This step typically involves writing code in the notebook to transform the raw data into a format suitable for BigQuery.
    - Perform the necessary data cleaning, manipulation, and preparation using Python and PySpark in the notebook.
4. Storing transformed data in Google Cloud Storage:
    - Once the data is transformed in the notebook, it needs to be stored back in GCS.
    - Ensure that the transformed data is saved in a specific location in GCS, ready to be loaded into BigQuery.
5. Loading data into BigQuery:

- Use commands or code to load the transformed data from GCS into BigQuery.
- Confirm that the data is successfully loaded into the desired tables in BigQuery.

Throughout the process, make sure to document the code, configurations, and steps taken in each phase of the ETL process. This documentation will help others understand and replicate the workflow.

Notes

1. Service Account Privileges:
    - Make sure the service account used for accessing the Drive API and transferring data to GCS has sufficient privileges.
    - Grant the necessary permissions to the service account to read data from Google Drive and write data to GCS.
    - The required permissions may include accessing and managing files in Google Drive and interacting with GCS buckets.
2. User Account Privileges:
    - Ensure that the user account you are using to set up and configure the ETL pipeline has appropriate privileges in the Google Drive where the data is saved.
    - Verify that the user account has the necessary permissions to create and manage service accounts, enable APIs, and generate credentials.
3. Granting Permissions in Google Drive:
    - For the specific Google Drive folder or files from which you're extracting data, ensure that the service account and user account have the appropriate access permissions.
    - Grant read access to the service account so that it can retrieve the data from Google Drive.
    - If you need to write data back to Google Drive at any point, ensure that the service account or user account has write permissions as well.

    By ensuring that both service accounts and user accounts have the necessary privileges, you can minimize potential issues related to accessing data from Google

Drive and storing it in GCS. It's essential to review and update these privileges as needed to maintain data integrity and security.