

General Questions

1. What are the differences among the files? are they different set's of sessions, different days or the same day split in 3 files?
2. Do we need the empty columns: jk, fgh, jyv? If so, do we expect some type of value
3. How frequently does the ETL process need to run? Is it a daily batch process or more frequent?
4. Are there any specific performance requirements, such as time limits within which the ETL process should complete?

Validation

5. Specific validations for the email field
Format: Can we safely assume it should be alphanumeric + period (.) as in Gmail.
Domain restrictions: Any domain restriction? If so, should it be whitelisting or blacklisting
6. Are there any specific error codes or messages that should be logged in the 'errors' table?

Database

I am assuming we use the email as ID for the user

7. For the `visitor` table, how should the ETL process update email entries? Depends on Question 1
Should it update the existing records based on the latest file, daily, or are there other rules?

File Management

8. For the backup, is there a preferred compression format for the zip files?
 1. Should I store them individually or in groups under certain rules.
 2. Naming convention

Execution Environment

9. Are there any access restrictions or firewall rules that need to be considered for connecting to the remote server (8.8.8.8)?

Error Handling and Notifications

10. What is the preferred method for error notifications? (e.g., email alerts, log files)
11. Are there any rollback mechanisms to consider in case the ETL process fails?

Additional Guidelines

12. Are there any specific coding languages or frameworks you prefer for implementing this ETL process?