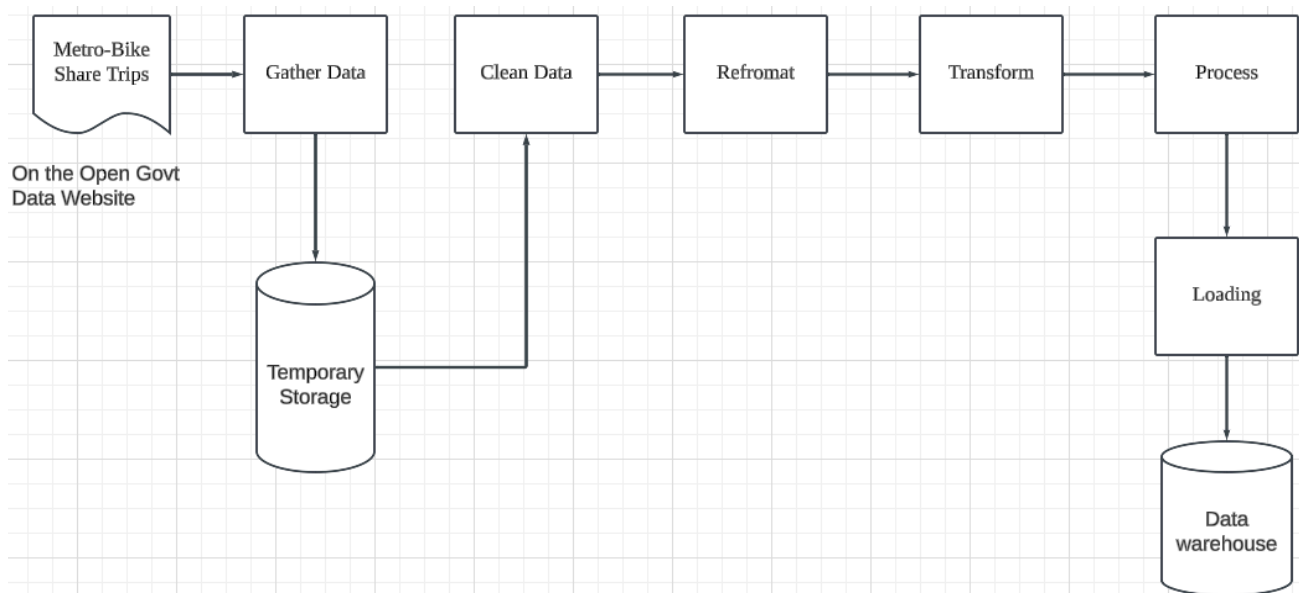


## Information Architecture



The data used in this research comes from a website designed to retrieve unprocessed data on different travels. To make sure the process runs well, such data is first removed and momentarily stored in a staging area. Because it serves as a buffer and guards against data loss or collision in the event of unexpected pauses or failures during the extraction process, this type of temporary storage is crucial. To make sure that the extracted data is dependable and not collide, the replication layer as a temporary storage can support the data processing pipeline in a simultaneous way and quick recovery in the event of problems. To make sure accuracy and error-free results, the captured data is getcleaned. Cleaning is an important process since it fixes errors and improves the quality. In summary, the procedure includes the following tasks: imputation of missing values, format standardization, eliminating duplicates, and data corrections. After cleaning, the data is reshaped and changed to make it feasible. Finally, the structured dataset is put into a data warehouse. At this level, the trip data is separated into fact and dimension tables in a star schema architecture to enable effective, scalable analytics. With the help of this schema, users may query and examine data on various time, location, and trip-related variables to gain insightful knowledge about things like popular routes, fare patterns, peak usage periods, and much more. From extraction to loading, our data pipeline guarantees that the final dataset is correct and comprehensive, allowing it to be pre-loaded for additional analysis.