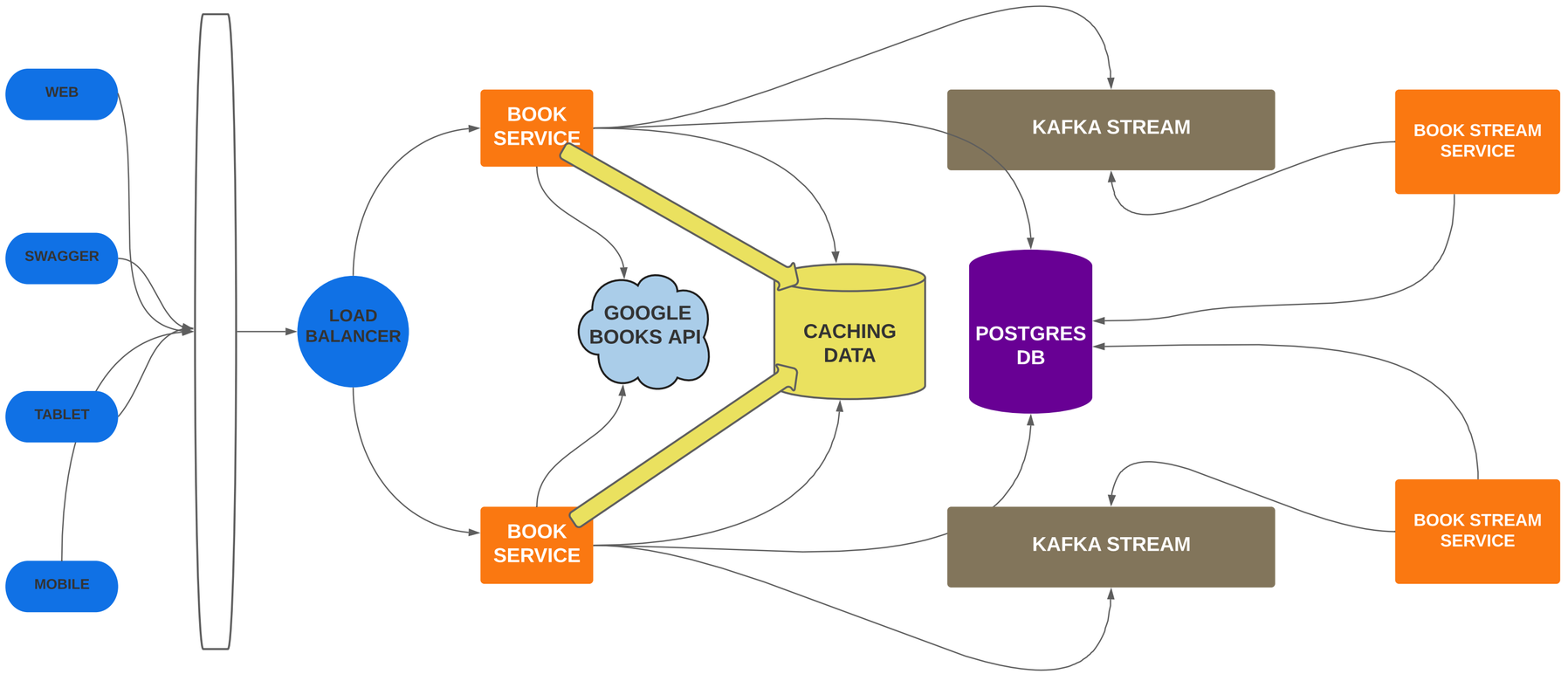

Load balancers have been used to scale system to handle more API requests.
--> LEAST LOADED mechanism will be used to distribute requests.
Book service has been scaled horizontally and the load balancer distributes API requests among services.
Kafka streams and book stream service have been scaled horizontally to handle more data processing and reading.
Since this is a read-heavy system, database can be replicated to handle more read operations.
Caching mechanism have been introduced to handle more read operations and reduce load on database.
--> CASH ASIDE strategy will be implemented since write operations will be less.
--> LEAST FREQUENTLY USED will be used as an cache eviction mechanish.