Rate limiter high level design

Limit request to server: 3 request per hour/minute/second

Functional requirements:

* The function will be run automatically before execution of the request function.
* The function doesn't overwhelm the system's capacity.

Non-functional requirements:

* The rate limiter will not affect the system if it isn’t available.
* The rate limiter always runs rightly whether the system uses sticky sessions or not.

In fact, we also need a place to store the rate limiter rules configured in yaml files. For simplicity, I just use one rule so we aren't necessary for placing a cached rules storage.

