1. Есть смысл не отдавать всю информацию для юзера через один API запрос, так как он может быть "тяжелым" (во время запроса к БД, используются сложные группировки и тд), а создать несколько ендпоинтов, и отдавать юзеру ответы ассинхронно для каждого запроса. тем самым клиент быстрее увидит информацию, так как каждый запрос будет выполняться меньше времени чем 1 большой (в одном выполнится обычный селект из БД, в другом выполнится группировка в БД и тд), и запросы выполнятся одновременно из за того что они ассинхронные
2. Добавить фильтры (тогда будет делаться запрос /api/v1/users/info?status=...&something=... - тем самым и в запрос к БД будет делаться фильтр WHERE, который вернет меньше записей и скорость ответа будет выше)
3. Добавить limit (/api/v1/users/info?limit=50) - можно реализовать в виде пагинации или подгрузки элементов при скроле вниз. Что бы знать от какого элемента отсчитывать, то так же можно применять limit=50&page=N или limit=50&offset=0 (значит с 1 элемента), limit=50&offset=50 (значит с 50 элемента)
4. Добавить REDIS cache
5. Добавить индексы в БД
6. Шардировать БД