Batch Job or One by One

Today, in our apartment, the water supply had some problems and couldn' t provide water for a while. After finishing the meal and looking at the piles of dishes that my family members and I had left, I thought of several questions.

One is how to let the washing machine continue working without a water supply. It may be designed to be connected to a bucket of water. Additionally, the attaching head of the water pipe should be flexible to easily switch from the public water supply to a private, self-made bucket of water.

Another point is the question of whether to do the job in batches or one by one. We can wash the dishes after every meal or wait to wash them after a day or a few days.

It reminds me of programming. We can perform tasks in batches or one by one.

Doing the job in batches leads to one apparent problem: it requires more resources. It needs more dishes as we delay the washing, and it requires more memory space as we accumulate the data to defer handling.

In real life, there is a limit to how much space or how many items can be handled at once. For example, the washing machine can handle up to twenty dishes at most, much like how a program has a memory limit on a computer or how a road has a limit on the number of cars that can pass through.

There is also the problem of how to separate the job. Should we separate it one item at a time or three items at a time?

For dishes or cars, it is simple to treat each item as a unit. That means a dish is a dish, and a car is a car. Normally, they cannot be broken down into smaller pieces. Though there are still exceptions, like a big truck that carries many cars; a big truck can be broken down into one large unit and many cars passing through the road.

In programming, it is much more flexible. Even an insert or update SQL can be broken down into smaller pieces, not to mention a download job, a DFS search, or a query.

OK, now we have thought about the handling unit. Then, the question is how many units should we process in one batch. It can be any number between one and the total number of units.

The question here is whether the number of batches for a job can be fixed or dynamic. For generative AI, the total characters of input text are flexible. It has some context limits or input limits, but within its limit range, it is flexible.

When using the washing machine, its inner space has a limit. Within that limit, the number of dishes is flexible. We normally put as many dishes that need washing into the machine.

For programs, the batch of SQLs for the database to handle at one time has a limit. Within that limit, the number of SQLs it can handle is flexible. But we should consider the network task of passing the SQLs from

the client to the database server, how much time the user can wait, and what happens if one unit task of the batch fails.

So for the problem of how many unit tasks we should do in one batch, we should consider the goal of the job, the limit of the downstream consumer or handler, and the probability of failure.

This way of thinking can be applied to many things. There are basically two problems to consider: what is the unit task and how many units should we process in a batch. By considering these problems, we may arrive at an optimal solution.