Trabajo por Lotes o Uno por Uno

Hoy, en nuestro apartamento, hubo algunos problemas con el suministro de agua y no pudimos tener agua durante un rato. Después de terminar de comer y ver la pila de platos que mis familiares y yo habíamos dejado, se me ocurrieron varias preguntas.

Uno es cómo hacer que el lavavajillas siga funcionando sin un suministro de agua. Podría estar diseñado para conectarse a un balde de agua. Además, la cabeza de conexión de la tubería de agua debe ser flexible para poder cambiar fácilmente del suministro público de agua a un balde de agua privado y hecho por uno mismo.

Otro punto es la cuestión de si hacer el trabajo por lotes o uno por uno. Podemos lavar los platos después de cada comida o esperar a lavarlos después de un día o unos días. Esto es desde el punto de vista de cómo comemos y lavamos los platos. También podemos abordar este problema desde el punto de vista de cuántos platos puede contener un lavavajillas.

Me recuerda a la programación. Podemos realizar tareas en lotes o una por una.

Hacer el trabajo en lotes conlleva un problema evidente: requiere más recursos. Se necesitan más platos al retrasar el lavado, y se requiere más espacio en la memoria al acumular los datos para posponer su procesamiento.

En la vida real, existe un límite en cuanto a la cantidad de espacio o elementos que se pueden manejar al mismo tiempo. Por ejemplo, el lavavajillas probablemente puede manejar un máximo de veinte platos, de manera similar a cómo un programa tiene un límite de memoria en una computadora o cómo una carretera tiene un límite en la cantidad de autos que pueden pasar por ella.

También está el problema de cómo separar el trabajo. ¿Deberíamos separarlo un elemento a la vez o tres elementos a la vez?

Para platos o coches, es sencillo tratar cada elemento como una unidad. Esto significa que un plato es un plato, y un coche es un coche. Normalmente, no se pueden descomponer en piezas más pequeñas. Aunque todavía hay excepciones, como un camión grande que transporta muchos coches; un camión grande se puede descomponer en una unidad grande y muchos coches que pasan por la carretera.

En programación, es mucho más flexible. Incluso una inserción o actualización SQL puede desglosarse en partes más pequeñas, sin mencionar un trabajo de descarga, una búsqueda DFS o una consulta.

OK, ahora hemos pensado en la unidad de manejo. Entonces, la pregunta es cuántas unidades

deberíamos procesar en un lote. Puede ser cualquier número entre uno y el número total de unidades.

La pregunta aquí es si el número de lotes para un trabajo puede ser fijo o dinámico. Para la IA generativa, la cantidad total de caracteres del texto de entrada es flexible. Tiene algunos límites de contexto o de entrada, pero dentro de su rango de límite, es flexible.

Al usar el lavavajillas, su espacio interior tiene un límite. Dentro de ese límite, la cantidad de platos es flexible. Normalmente colocamos tantos platos como sea necesario lavar en la máquina.

Para los programas, el lote de SQLs que la base de datos puede manejar a la vez tiene un límite. Dentro de ese límite, la cantidad de SQLs que puede manejar es flexible. Sin embargo, debemos considerar la tarea de red de pasar los SQLs desde el cliente al servidor de la base de datos, cuánto tiempo puede esperar el usuario y qué sucede si una tarea unitaria del lote falla.

Entonces, para el problema de cuántas tareas unitarias debemos realizar en un lote, debemos considerar el objetivo del trabajo, el límite del consumidor o manejador posterior, y la probabilidad de fallo.

Esta forma de pensar puede aplicarse a muchas cosas. Básicamente, hay dos problemas a considerar: cuál es la tarea unitaria y cuántas unidades debemos procesar en un lote. Al considerar estos problemas, podemos llegar a una solución óptima.