1. It helps identify potential bottlenecks; this is important if a system is expected to handle a certain number of concurrent users and can alert developers to optimize for better performance.
2. It can help in planning ahead, if developers are able to identify the systems bottleneck, they can be aware of a need to add more resources to scale the system.