






Performance and Scalability patterns

06/23/2017 • 2 minutes to read • Contributors     

Performance is an indication of the responsiveness of a system to execute any action within a given time interval, while scalability is ability of a system either to handle increases in load without impact on performance or for the available resources to be readily increased. Cloud applications typically encounter variable workloads and peaks in activity. Predicting these, especially in a multi-tenant scenario, is almost impossible. Instead, applications should be able to scale out within limits to meet peaks in demand, and scale in when demand decreases. Scalability concerns not just compute instances, but other elements such as data storage, messaging infrastructure, and more.

| Pattern | Summary |
|---------------------------|--|
| Cache-Aside | Load data on demand into a cache from a data store |
| CQRS | Segregate operations that read data from operations that update data by using separate interfaces. |
| Event Sourcing | Use an append-only store to record the full series of events that describe actions taken on data in a domain. |
| Index Table | Create indexes over the fields in data stores that are frequently referenced by queries. |
| Materialized View | Generate prepopulated views over the data in one or more data stores when the data isn't ideally formatted for required query operations. |
| Priority Queue | Prioritize requests sent to services so that requests with a higher priority are received and processed more quickly than those with a lower priority. |
| Queue-Based Load Leveling | Use a queue that acts as a buffer between a task and a service that it invokes in order to smooth intermittent heavy loads. |
| Sharding | Divide a data store into a set of horizontal partitions or shards. |
| Static Content Hosting | Deploy static content to a cloud-based storage service that can deliver them directly to the client. |
| Throttling | Control the consumption of resources used by an instance of an application, an individual tenant, or an entire service. |