Scaling Applications Using Windows Azure Cloud Services

A key benefit of Windows Azure is creating highly scalable applications using Cloud Services. Applications can shrink and stretch to accommodate changes in usage, removing the need for expensive on-premises hardware. A key strategy is to design in scale units, which are a base configuration of web and worker role instances with supporting services such as data stores and caching.

There are three reasons to create Windows Azure scalable applications:

1. **Demand Peaks**: Your application has thousands of users (or more) although usage varies, sometimes greatly.
2. **Distributed Users and Devices**: Your users are spread out, even around the globe.
3. **Partitionable Workloads**: Your workloads are distributed; optimal data loads of work, which can be scaled with Cloud Services.

**Plan and Design**

A highly scalable application requires the use of specific patterns and practices. Outlining for optimal performance and scale out guide. Use the patterns below to help you architect your solution and continually refine your application.

- **Scale Out with Scale Units**
- **Decoupled Communications**
- **Betty for Fault Tolerance**
- **Caching**
- **Horizontal Partitioning**
- **Vertical Affinity**

**Build and Deploy**

Cloud Services are built for scalability. Web and worker instances can be increased and decreased at will. Workloads can be distributed using messaging, such as queues or Service Bus Topics.

**Run and Tune**

This phase contains the processes that refine the application, keep it running, and make it perform. Automation allows you to remove the manual tasks and focus on monitoring. It is a good practice to continually assess the metrics and balance against running costs.