**Rate Limiting Pattern**

To prevent excessive resource consumption, some services impose limits on how fast other applications or services can access them. This is called throttling. You can use a rate limiting pattern to reduce or avoid throttling errors caused by these limits and to estimate throughput more accurately.

A rate-limiting pattern can be useful in many situations, but it is especially helpful for large-scale repetitive automated tasks such as batch processing.

By limiting the amount of records supplied to a service during a specific time, rate limiting can lower your traffic and possibly increase throughput.

Various measures may be used to throttle a service over time, including:

— the quantity of actions (60 requests, for instance).  
— the volume of data (50 GB per minute, for instance).  
— the relative expense of operations (42,000 RUs per second, for instance).

Regardless of the metric employed for throttling, the rate limiting approach you choose will include regulating the volume and/or size of operations delivered to the service over a predetermined time frame in order to maximize your use of the service without going over the throttling limit.