100 XP

Discover request units

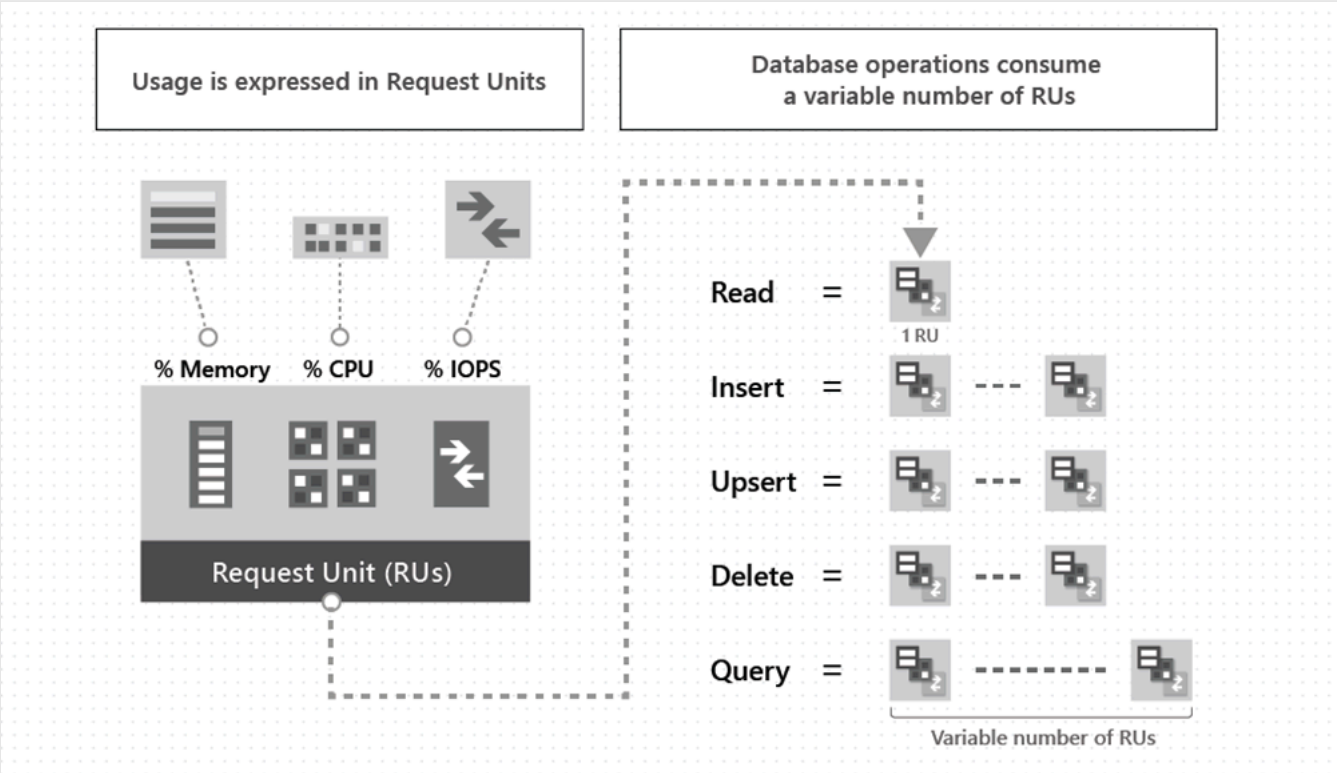
3 minutes

With Azure Cosmos DB, you pay for the throughput you provision and the storage you consume on an hourly basis. Throughput must be provisioned to ensure that sufficient system resources are available for your Azure Cosmos database always.

The cost of all database operations is normalized by Azure Cosmos DB and is expressed by *request units* (or RUs, for short). A request unit represents the system resources such as CPU, IOPS, and memory that are required to perform the database operations supported by Azure Cosmos DB.

The cost to do a point read, which is fetching a single item by its ID and partition key value, for a 1-KB item is 1RU. All other database operations are similarly assigned a cost using RUs. No matter which API you use to interact with your Azure Cosmos container, costs are measured by RUs. Whether the database operation is a write, point read, or query, costs are measured in RUs.

The following image shows the high-level idea of RUs:



The type of Azure Cosmos DB account you're using determines the way consumed RUs get charged. There are three modes in which you can create an account:

- **Provisioned throughput mode**: In this mode, you provision the number of RUs for your application on a per-second basis in increments of 100 RUs per second. To scale the provisioned throughput for your application, you can increase or decrease the number of RUs at any time in increments or decrements of 100 RUs. You can make your changes either programmatically or by using the Azure portal. You can provision throughput at container and database granularity level.
- **Serverless mode**: In this mode, you don't have to provision any throughput when creating resources in your Azure Cosmos DB account. At the end of your billing period, you get billed for the number of request units that have been consumed by your database operations.
- **Autoscale mode:** In this mode, you can automatically and instantly scale the throughput (RU/s) of your database or container based on its usage. This scaling operation doesn't affect the availability, latency, throughput, or performance of the workload. This mode is well suited for mission-critical workloads that have variable or unpredictable traffic patterns, and require SLAs on high performance and scale.

Next unit: Exercise: Create Azure Cosmos DB resources by using the Azure portal

Continue >