Tuning for Throughput and Performance

Leonard Lobel
CTO, SLEEK TECHNOLOGIES

lennilobel.wordpress.com

Measuring Performance

Latency

How fast is the response for a given request?

Throughput

How many requests can be served within a specific period of time?

Introducing Request Units

Throughput Currency

Blended measure of computational cost (CPU, memory, disk I/O, network I/O)

All Requests are Not Equal

Every Cosmos DB response header shows the RU charge for the request

Request Units are Deterministic

The same request will always require the same number of request units

Reserving Request Units

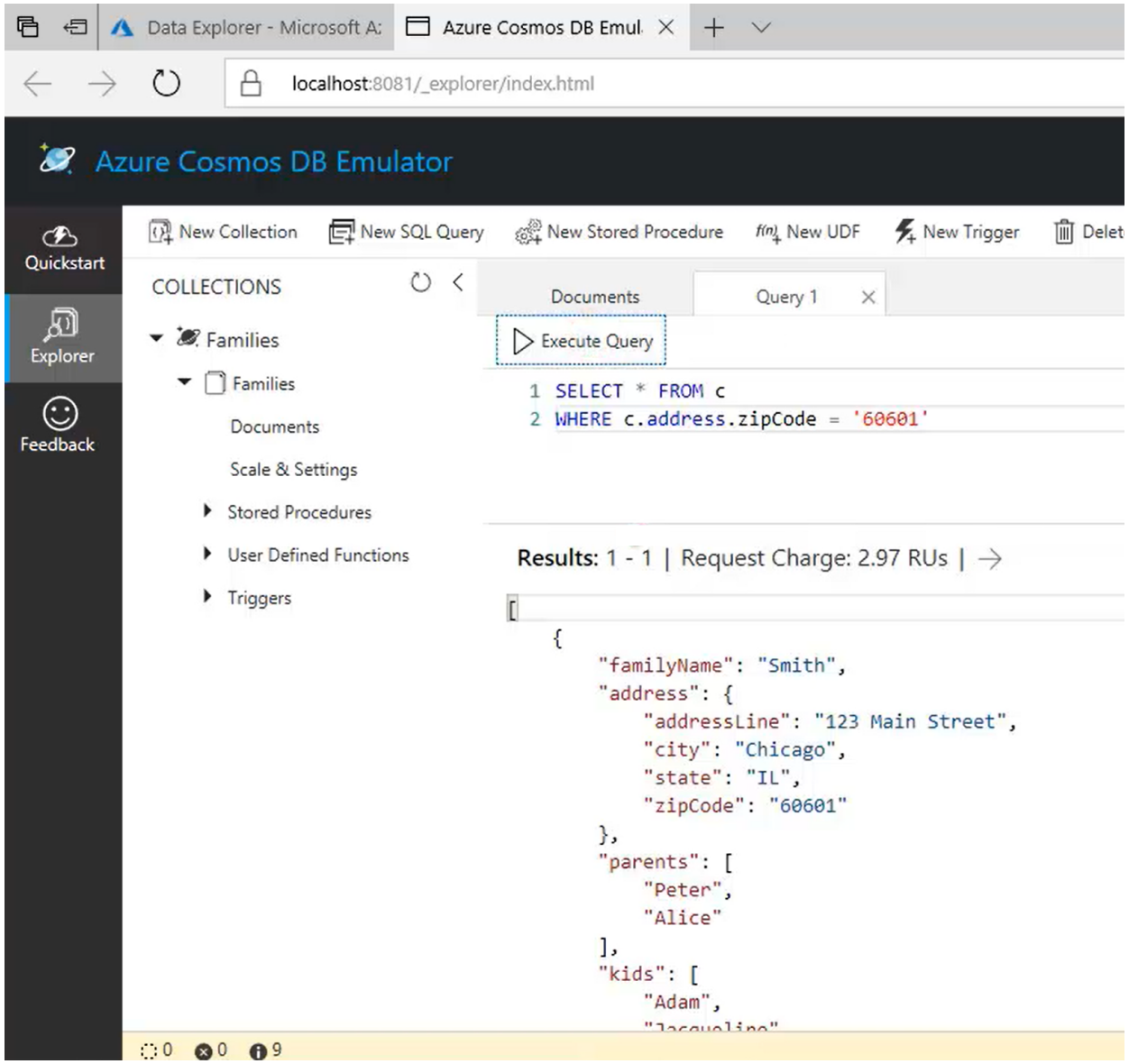
Provision request units per second (RU/s)

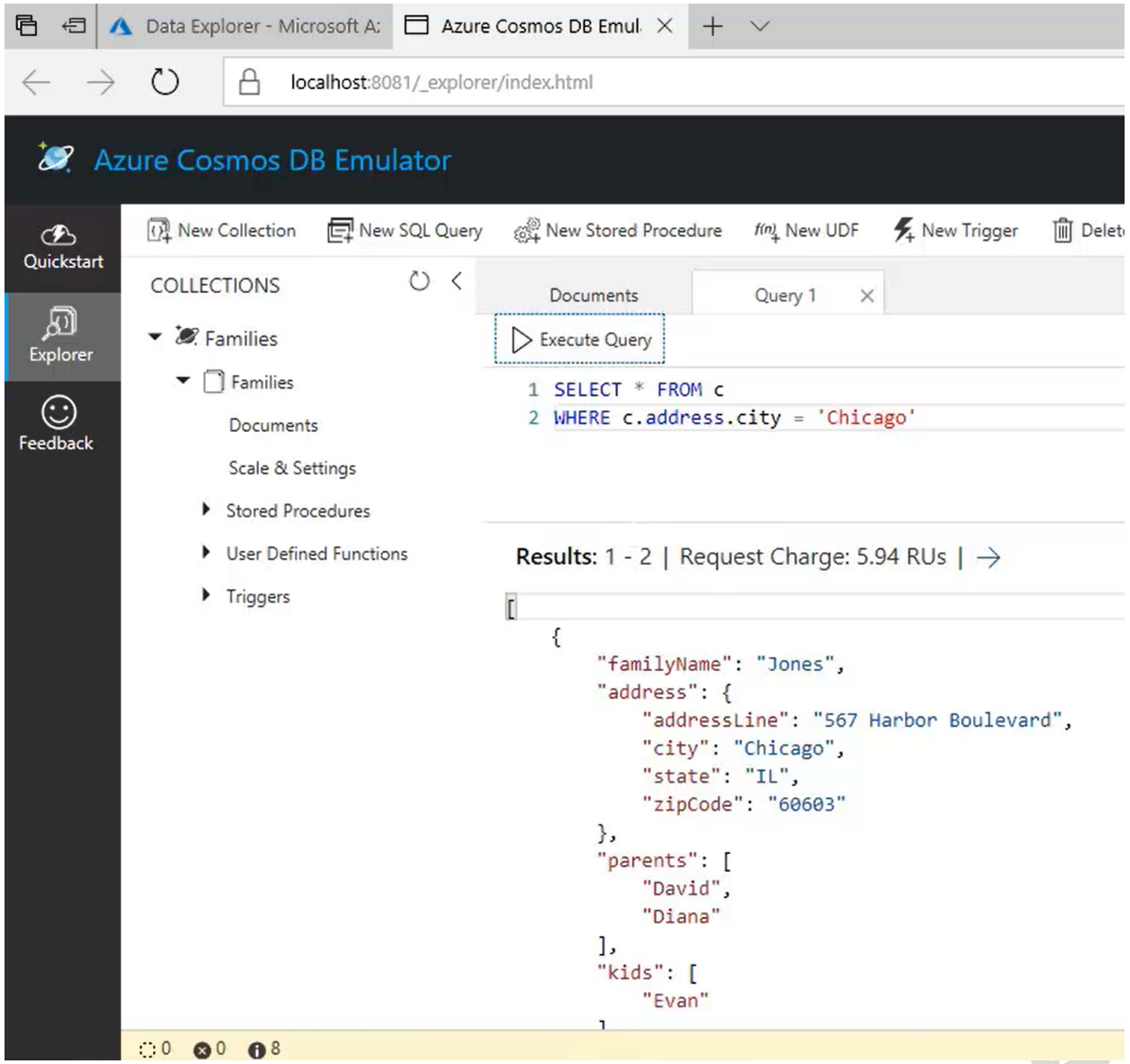
How many request *units* (not *requests*) per second are available to your application

Exceeding reserved throughput limits

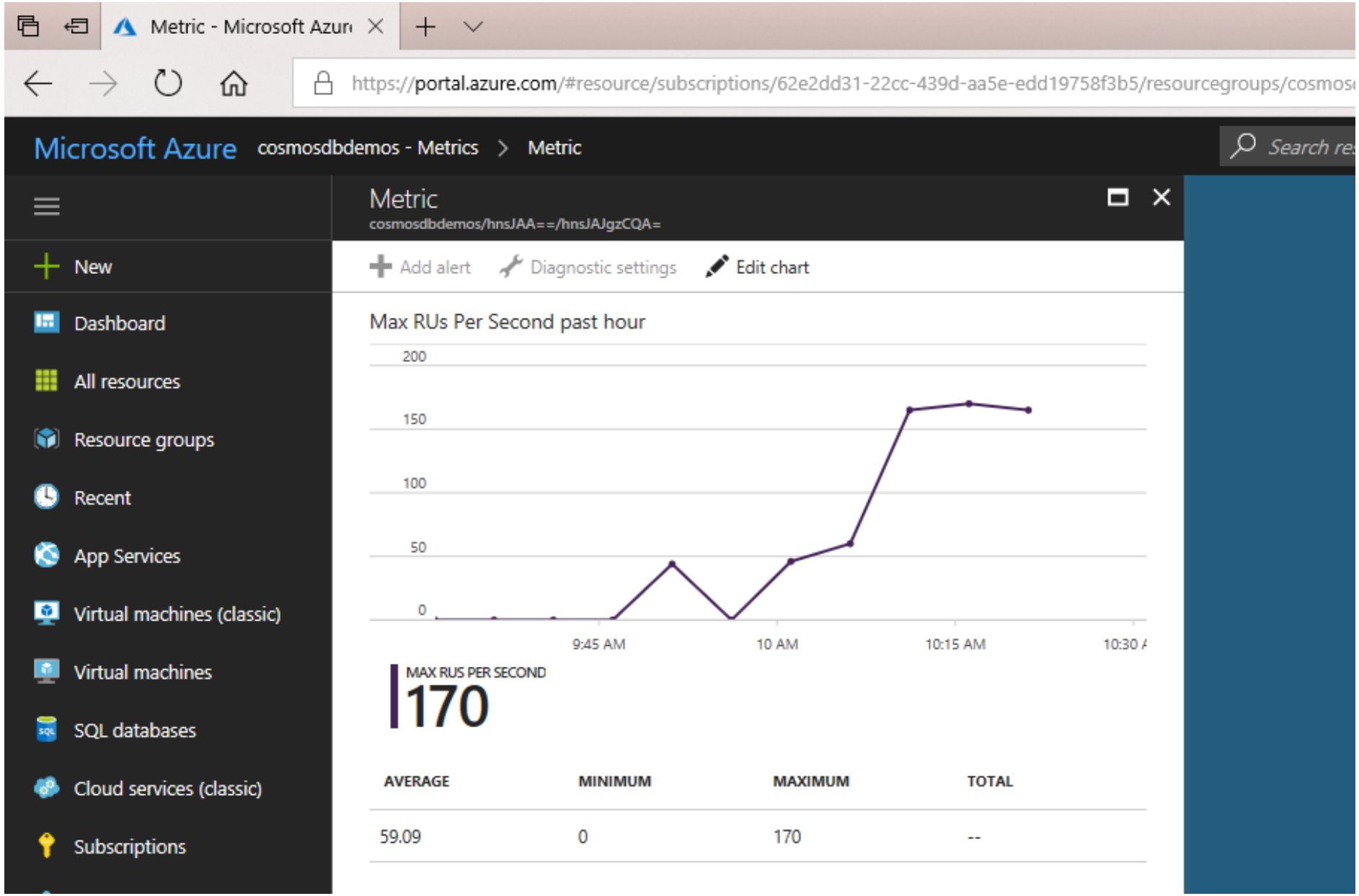
Requests are “throttled”

Monitoring Request Unit Consumption





Monitoring Request Unit Consumption



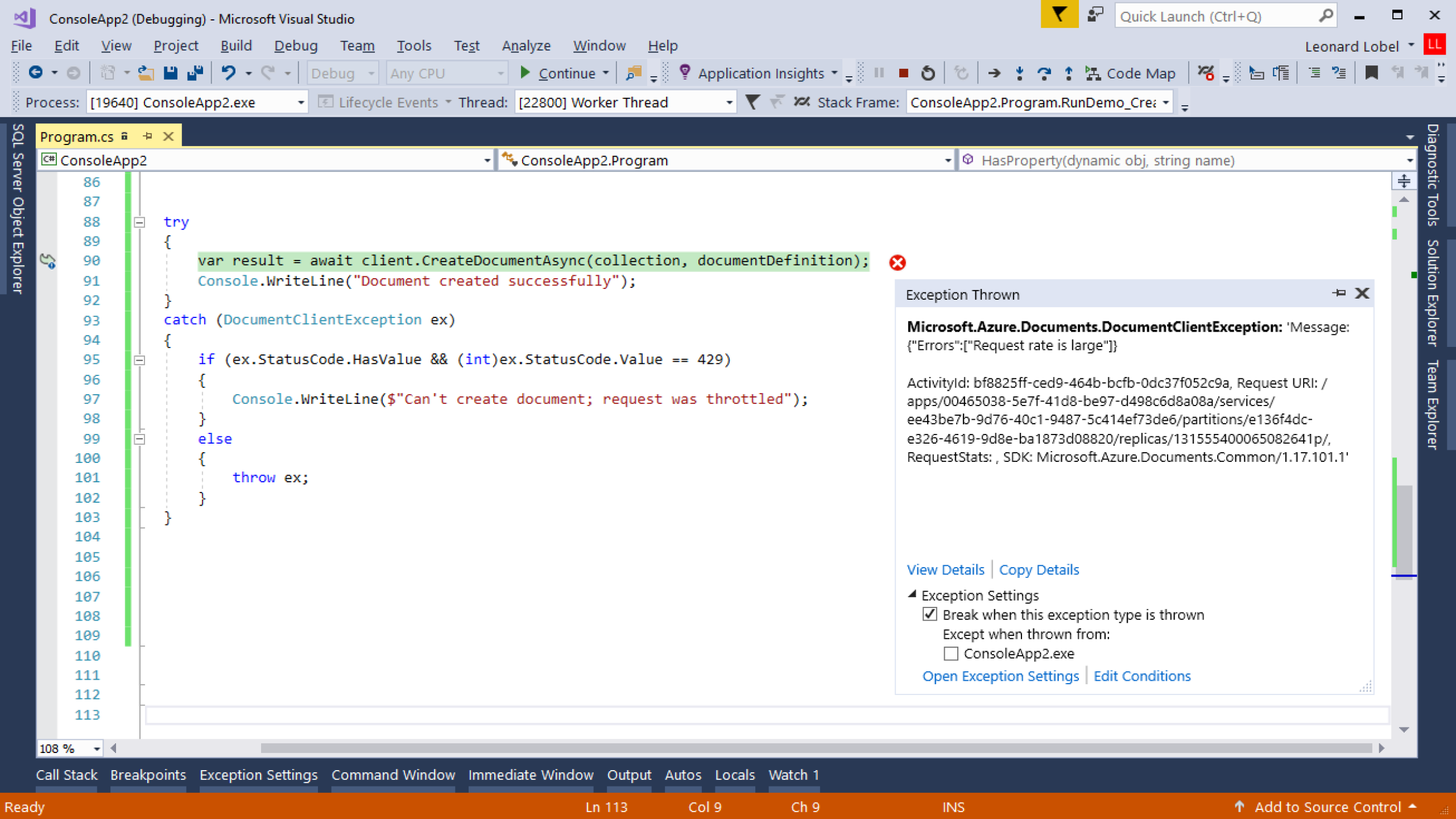
Monitoring Request Unit Consumption

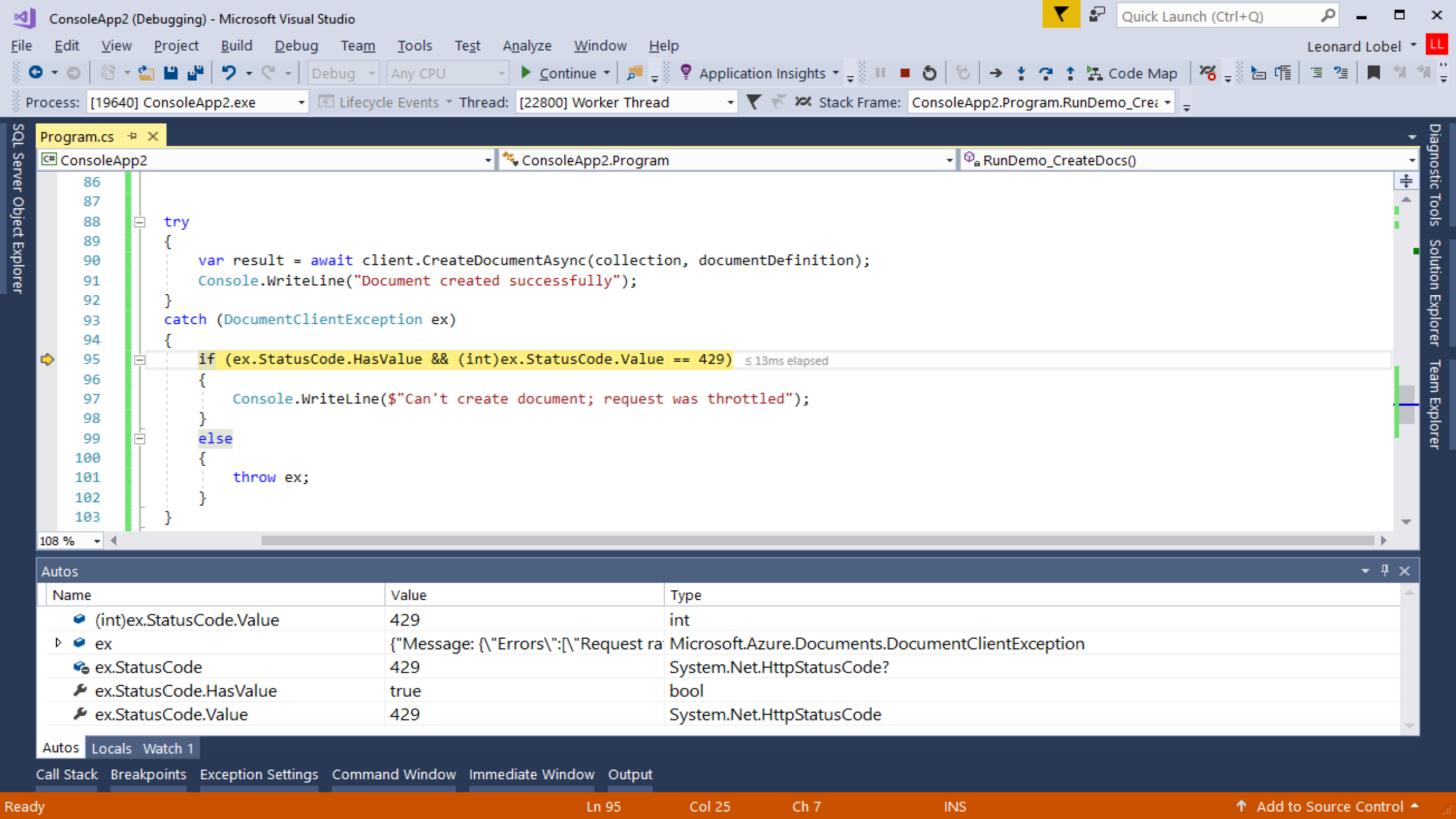
Program.cs

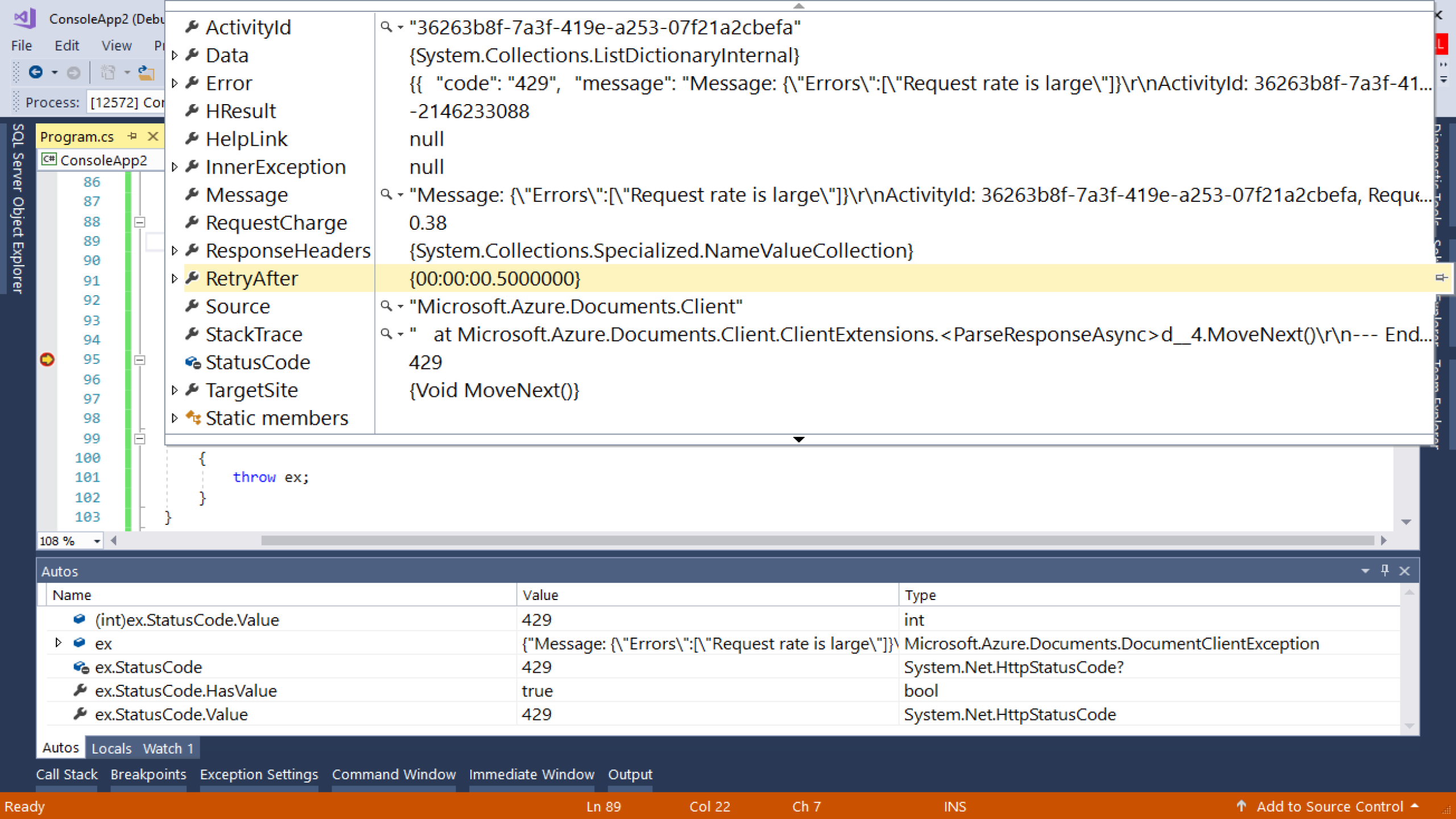
ConsoleApp2 | ConsoleApp2.Program | RunDemo_CreateDocs()

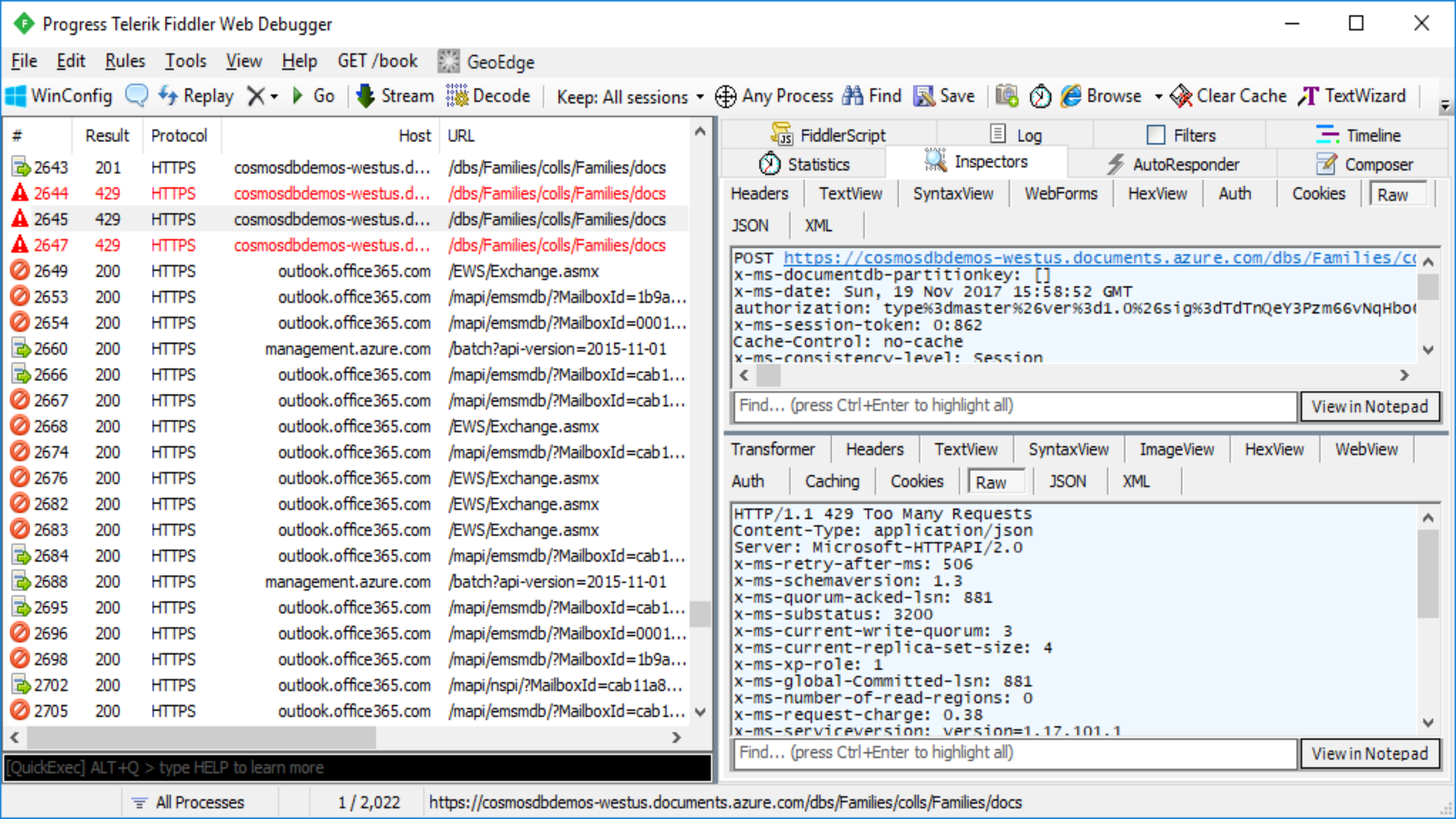
```
53
54      var collection = UriFactory.CreateDocumentCollectionUri("Families", "Families");
55
56      dynamic documentDefinition = new
57      {
58          familyName = "Smith",
59          address = new
60          {
61              addressLine = "123 Main Street",
62              city = "Chicago",
63              state = "IL",
64              zipCode = "60601"
65          },
66          parents = new string[]
67          {
68              "Peter",
69              "Alice"
70          },
71          kids = new string[]
72          {
73              "Adam",
74              "Jacqueline",
75              "Joshua"
76          },
77      };
78
79      var result = await client.CreateDocumentAsync(collection, documentDefinition);
80      var consumedRUs = result.RequestCharge;
81
82      Console.WriteLine($"Cost to create document: {consumedRUs} RUs");  ≤ 194ms elapsed
83                                                                          consumedRUs 8.95
84
85
```

Exceeding Reserved Throughput Limits


ConsoleApp2 (Debugging) - Microsoft Visual Studio
Quick Launch (Ctrl+Q)
File Edit View Project Build Debug Team Tools Test Analyze Window Help
Leonard Lobel
Debug
Any CPU
Continue
Application Insights
Code Map
Process: [19640] ConsoleApp2.exe
Lifecycle Events
Thread: [22800] Worker Thread
Stack Frame: ConsoleApp2.Program.RunDemo_Crea
SQL Server Object Explorer
Program.cs
ConsoleApp2
ConsoleApp2.Program
HasProperty(dynamic obj, string name)
86
87
88 try
89 {
90 var result = await client.CreateDocumentAsync(collection, documentDefinition);
91 Console.WriteLine("Document created successfully");
92 }
93 catch (DocumentClientException ex)
94 {
95 if (ex.StatusCode.HasValue && (int)ex.StatusCode.Value == 429)
96 {
97 Console.WriteLine($"Can't create document; request was throttled");
98 }
99 else
100 {
101 throw ex;
102 }
103 }
104
105
106
107
108
109
110
111
112
113
Exception Thrown
Microsoft.Azure.Documents.DocumentClientException: 'Message: {"Errors":["Request rate is large"]}
ActivityId: bf8825ff-ced9-464b-bcfb-0dc37f052c9a, Request URI: /apps/00465038-5e7f-41d8-be97-d498c6d8a08a/services/ee43be7b-9d76-40c1-9487-5c414ef73de6/partitions/e136f4dc-e326-4619-9d8e-ba1873d08820/replicas/131555400065082641p/, RequestStats: , SDK: Microsoft.Azure.Documents.Common/1.17.101.1'
View Details
Copy Details
Exception Settings
Break when this exception type is thrown
Except when thrown from:
ConsoleApp2.exe
Open Exception Settings
Edit Conditions
108 %
Call Stack Breakpoints Exception Settings Command Window Immediate Window Output Autos Locals Watch 1
Diagnostic Tools Solution Explorer Team Explorer
Ready
Ln 113 Col 9 Ch 9 INS
Add to Source Control


ConsoleApp2 (Debugging) - Microsoft Visual Studio
Quick Launch (Ctrl+Q)
File Edit View Project Build Debug Team Tools Test Analyze Window Help
Leonard Lobel
Debug
Any CPU
Continue
Application Insights
Code Map
Process: [19640] ConsoleApp2.exe
Lifecycle Events
Thread: [22800] Worker Thread
Stack Frame: ConsoleApp2.Program.RunDemo_Crea
SQL Server Object Explorer
Program.cs
ConsoleApp2
ConsoleApp2.Program
RunDemo_CreateDocs()
86
87
88 try
89 {
90 var result = await client.CreateDocumentAsync(collection, documentDefinition);
91 Console.WriteLine("Document created successfully");
92 }
93 catch (DocumentClientException ex)
94 {
95 if (ex.StatusCode.HasValue && (int)ex.StatusCode.Value == 429) ≤ 13ms elapsed
96 {
97 Console.WriteLine($"Can't create document; request was throttled");
98 }
99 else
100 {
101 throw ex;
102 }
103 }
108 %
Autos
Name
Value
Type
(int)ex.StatusCode.Value
429
int
ex
{"Message: {\"Errors\":[\"Request ra
Microsoft.Azure.Documents.DocumentClientException
ex.StatusCode
429
System.Net.HttpStatusCode?
ex.StatusCode.HasValue
true
bool
ex.StatusCode.Value
429
System.Net.HttpStatusCode
Autos
Locals
Watch 1
Call Stack
Breakpoints
Exception Settings
Command Window
Immediate Window
Output
Ready
Ln 95
Col 25
Ch 7
INS
Add to Source Control
Diagnostic Tools
Solution Explorer
Team Explorer


ConsoleApp2 (Debu
File Edit View Pr
Process: [12572] Cor
SQL Server Object Explorer
Program.cs
ConsoleApp2
ActivityId "36263b8f-7a3f-419e-a253-07f21a2cbefa"
Data {System.Collections.ListDictionaryInternal}
Error {{ "code": "429", "message": "Message: {\"Errors\":[\"Request rate is large\"]}\r\nActivityId: 36263b8f-7a3f-41...
HResult -2146233088
HelpLink null
InnerException null
Message "Message: {\"Errors\":[\"Request rate is large\"]}\r\nActivityId: 36263b8f-7a3f-419e-a253-07f21a2cbefa, Reque...
RequestCharge 0.38
ResponseHeaders {System.Collections.Specialized.NameValueCollection}
RetryAfter {00:00:00.5000000}
Source "Microsoft.Azure.Documents.Client"
StackTrace " at Microsoft.Azure.Documents.Client.ClientExtensions.<ParseResponseAsync>d__4.MoveNext()\r\n--- End...
StatusCode 429
TargetSite {Void MoveNext()}
Static members
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100 {
101 throw ex;
102 }
103 }
108 %
Autos
Name
Value
Type
(int)ex.StatusCode.Value 429 int
ex {"Message: {\"Errors\":[\"Request rate is large\"]}\ Microsoft.Azure.Documents.DocumentClientException
ex.StatusCode 429 System.Net.HttpStatusCode?
ex.StatusCode.HasValue true bool
ex.StatusCode.Value 429 System.Net.HttpStatusCode
Autos Locals Watch 1
Call Stack Breakpoints Exception Settings Command Window Immediate Window Output
Ready
Ln 89 Col 22 Ch 7 INS
Add to Source Control


Progress Telerik Fiddler Web Debugger
File Edit Rules Tools View Help GET /book GeoEdge
WinConfig Replay Go Stream Decode Keep: All sessions Any Process Find Save Browse Clear Cache TextWizard
Result Protocol Host URL
2643 201 HTTPS cosmosdbdemos-westus.d... /dbs/Families/colls/Families/docs
2644 429 HTTPS cosmosdbdemos-westus.d... /dbs/Families/colls/Families/docs
2645 429 HTTPS cosmosdbdemos-westus.d... /dbs/Families/colls/Families/docs
2647 429 HTTPS cosmosdbdemos-westus.d... /dbs/Families/colls/Families/docs
2649 200 HTTPS outlook.office365.com /EWS/Exchange.asmx
2653 200 HTTPS outlook.office365.com /mapi/emsmdb/?MailboxId=1b9a...
2654 200 HTTPS outlook.office365.com /mapi/emsmdb/?MailboxId=0001...
2660 200 HTTPS management.azure.com /batch?api-version=2015-11-01
2666 200 HTTPS outlook.office365.com /mapi/emsmdb/?MailboxId=cab1...
2667 200 HTTPS outlook.office365.com /mapi/emsmdb/?MailboxId=cab1...
2668 200 HTTPS outlook.office365.com /EWS/Exchange.asmx
2674 200 HTTPS outlook.office365.com /mapi/emsmdb/?MailboxId=cab1...
2676 200 HTTPS outlook.office365.com /EWS/Exchange.asmx
2682 200 HTTPS outlook.office365.com /EWS/Exchange.asmx
2683 200 HTTPS outlook.office365.com /EWS/Exchange.asmx
2684 200 HTTPS outlook.office365.com /mapi/emsmdb/?MailboxId=cab1...
2688 200 HTTPS management.azure.com /batch?api-version=2015-11-01
2695 200 HTTPS outlook.office365.com /mapi/emsmdb/?MailboxId=cab1...
2696 200 HTTPS outlook.office365.com /mapi/emsmdb/?MailboxId=0001...
2698 200 HTTPS outlook.office365.com /mapi/emsmdb/?MailboxId=1b9a...
2702 200 HTTPS outlook.office365.com /mapi/nspi/?MailboxId=cab11a8...
2705 200 HTTPS outlook.office365.com /mapi/emsmdb/?MailboxId=cab1...
[QuickExec] ALT+Q > type HELP to learn more
FiddlerScript Log Filters Timeline
Statistics Inspectors AutoResponder Composer
Headers TextView SyntaxView WebForms HexView Auth Cookies Raw
JSON XML
POST https://cosmosdbdemos-westus.documents.azure.com/dbs/Families/c
x-ms-documentdb-partitionkey: []
x-ms-date: Sun, 19 Nov 2017 15:58:52 GMT
authorization: type%3dmaster%26ver%3d1.0%26sig%3dTdTnQeY3Pzm66vNqHbo
x-ms-session-token: 0:862
Cache-Control: no-cache
x-ms-consistency-level: Session
Find... (press Ctrl+Enter to highlight all)
View in Notepad
Transformer Headers TextView SyntaxView ImageView HexView WebView
Auth Caching Cookies Raw JSON XML
HTTP/1.1 429 Too Many Requests
Content-Type: application/json
Server: Microsoft-HTTPAPI/2.0
x-ms-retry-after-ms: 506
x-ms-schemaversion: 1.3
x-ms-quorum-acked-lsn: 881
x-ms-substatus: 3200
x-ms-current-write-quorum: 3
x-ms-current-replica-set-size: 4
x-ms-xp-role: 1
x-ms-global-Committed-lsn: 881
x-ms-number-of-read-regions: 0
x-ms-request-charge: 0.38
x-ms-serviceversion: version=1.17.101.1
Find... (press Ctrl+Enter to highlight all)
View in Notepad
All Processes 1 / 2,022 https://cosmosdbdemos-westus.documents.azure.com/dbs/Families/colls/Families/docs

Exceeding Reserved Throughput Limits

Whiteboarding the Cost

Application checklist

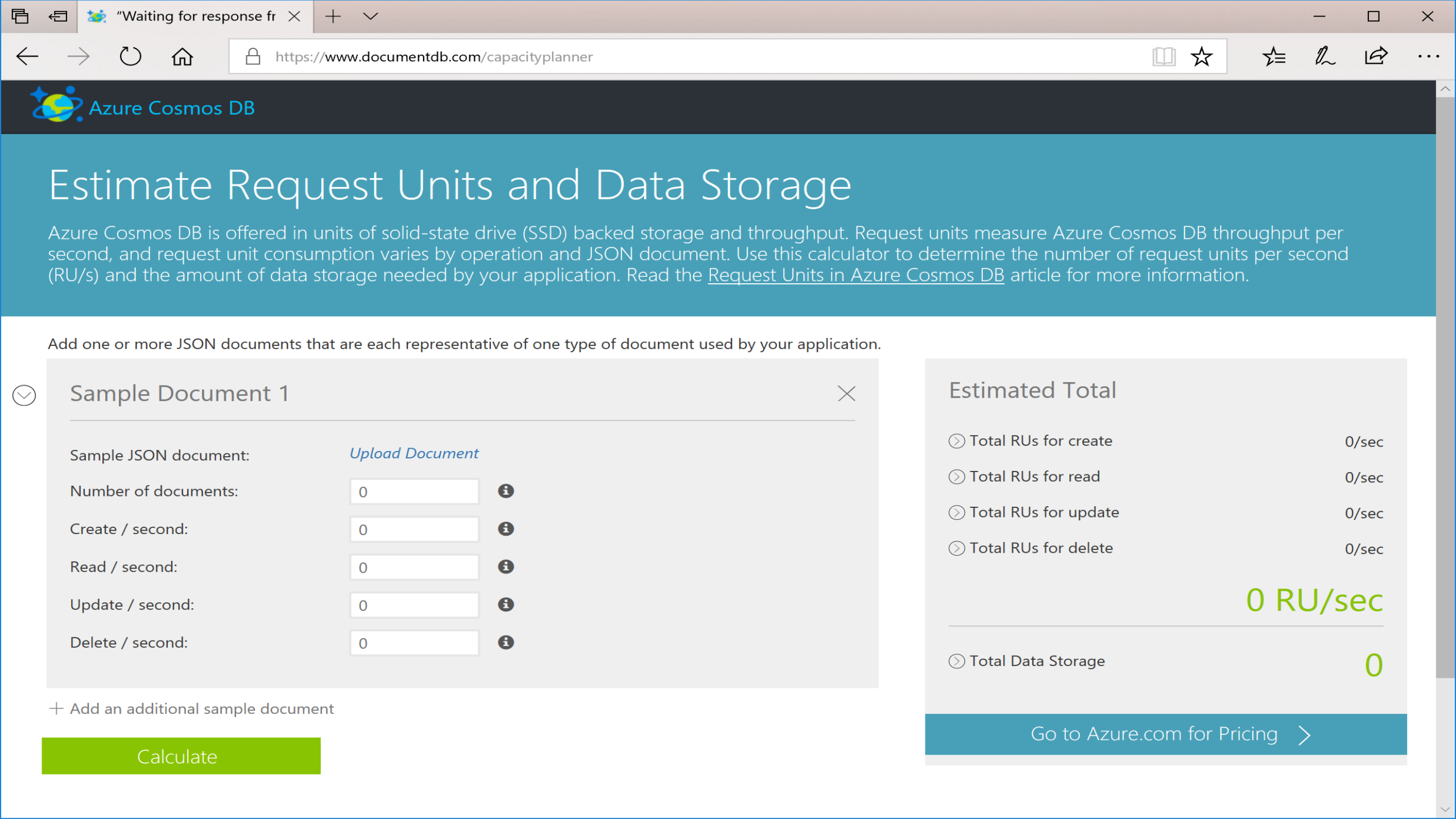
- What does a typical item look like?
- What are the typical queries that users will run?
- How many writes per second are required?
- How many queries per second are required?
- What is the acceptable consistency level?
- What is the indexing policy?

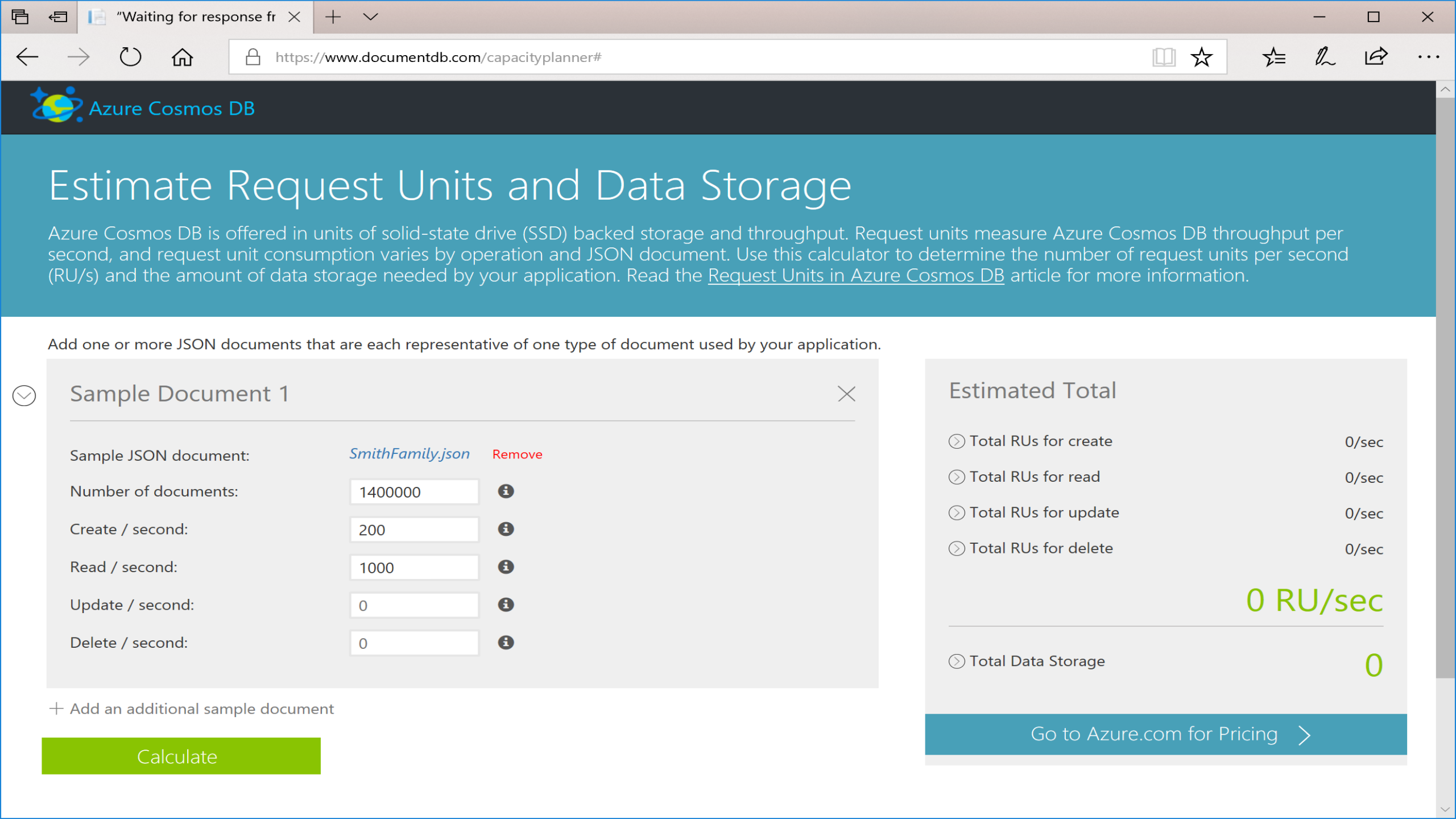
Estimating throughput needs

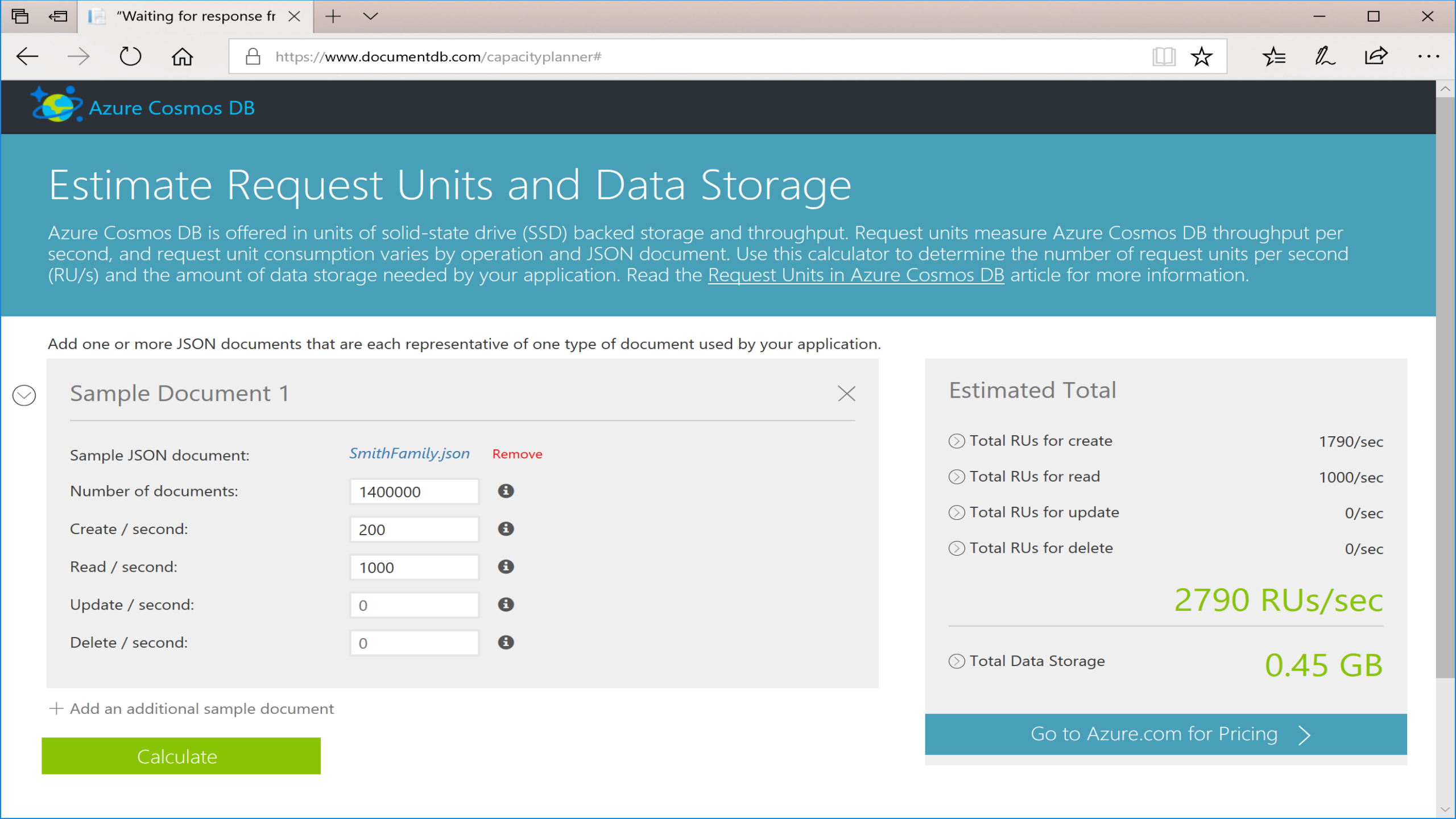
Item size	Reads/second	Writes/second	Request units
1 KB	500	100	(500 * 1) + (100 * 5) = 1,000 RU/s
1 KB	500	500	(500 * 1) + (500 * 5) = 3,000 RU/s
4 KB	500	100	(500 * 1.3) + (100 * 7) = 1,350 RU/s
4 KB	500	500	(500 * 1.3) + (500 * 7) = 4,150 RU/s
64 KB	500	100	(500 * 10) + (100 * 48) = 9,800 RU/s
64 KB	500	500	(500 * 10) + (500 * 48) = 29,000 RU/s

* Based on Session consistency indexing policy set to None.

Request Unit Calculator


"Waiting for response fr
https://www.documentdb.com/capacityplanner
Azure Cosmos DB
Estimate Request Units and Data Storage
Azure Cosmos DB is offered in units of solid-state drive (SSD) backed storage and throughput. Request units measure Azure Cosmos DB throughput per second, and request unit consumption varies by operation and JSON document. Use this calculator to determine the number of request units per second (RU/s) and the amount of data storage needed by your application. Read the Request Units in Azure Cosmos DB article for more information.
Add one or more JSON documents that are each representative of one type of document used by your application.
Sample Document 1
Sample JSON document:
Upload Document
Number of documents:
0
Create / second:
0
Read / second:
0
Update / second:
0
Delete / second:
0
Add an additional sample document
Calculate
Estimated Total
Total RUs for create
0/sec
Total RUs for read
0/sec
Total RUs for update
0/sec
Total RUs for delete
0/sec
0 RU/sec
Total Data Storage
0
Go to Azure.com for Pricing


"Waiting for response fr
https://www.documentdb.com/capacityplanner#
Azure Cosmos DB
Estimate Request Units and Data Storage
Azure Cosmos DB is offered in units of solid-state drive (SSD) backed storage and throughput. Request units measure Azure Cosmos DB throughput per second, and request unit consumption varies by operation and JSON document. Use this calculator to determine the number of request units per second (RU/s) and the amount of data storage needed by your application. Read the Request Units in Azure Cosmos DB article for more information.
Add one or more JSON documents that are each representative of one type of document used by your application.
Sample Document 1
Sample JSON document:
SmithFamily.json
Remove
Number of documents:
1400000
Create / second:
200
Read / second:
1000
Update / second:
0
Delete / second:
0
Add an additional sample document
Calculate
Estimated Total
Total RUs for create
0/sec
Total RUs for read
0/sec
Total RUs for update
0/sec
Total RUs for delete
0/sec
0 RU/sec
Total Data Storage
0
Go to Azure.com for Pricing


"Waiting for response fr
https://www.documentdb.com/capacityplanner#
Azure Cosmos DB
Estimate Request Units and Data Storage
Azure Cosmos DB is offered in units of solid-state drive (SSD) backed storage and throughput. Request units measure Azure Cosmos DB throughput per second, and request unit consumption varies by operation and JSON document. Use this calculator to determine the number of request units per second (RU/s) and the amount of data storage needed by your application. Read the Request Units in Azure Cosmos DB article for more information.
Add one or more JSON documents that are each representative of one type of document used by your application.
Sample Document 1
Sample JSON document:
SmithFamily.json
Remove
Number of documents:
1400000
Create / second:
200
Read / second:
1000
Update / second:
0
Delete / second:
0
Add an additional sample document
Calculate
Estimated Total
Total RUs for create
1790/sec
Total RUs for read
1000/sec
Total RUs for update
0/sec
Total RUs for delete
0/sec
2790 RUs/sec
Total Data Storage
0.45 GB
Go to Azure.com for Pricing

Pricing

SSD Storage

	A	B	C
1	1 GB	10 GB	
2	$ 0.25	$ 2.50	/month

Throughput

	A	B	C
1	100 RU/s	400 RU/s	
2	$ 0.008	$ 0.032	/hour
3	$ 0.192	$ 0.768	/day
4	$ 5.856	$ 23.424	/month

Summary

Measuring performance

- Latency and throughput

Request units

- Throughput currency
- Predictable throughput
- Monitoring consumption
- Calculating cost

Pricing

- Storage (consumption based)
- Throughput (reserved RU/s)