Microsoft

Azure
SQL Data Warehouse

Predictable performance with advanced workload management

Resource classes

Overview

Pre-determined resource limits defined for a user or role.

Govern the system memory assigned to each query.

Effectively used to control the number of concurrent queries that can run on a data warehouse.

```
/* View resource classes in the data warehouse */
SELECT name
FROM   sys.database_principals
WHERE  name LIKE '%rc%' AND type_desc = 'DATABASE_ROLE';

/* Change user’s resource class to 'largerc' */
EXEC sp_addrolemember 'largerc', 'loaduser’;

/* Decrease the loading user's resource class */
EXEC sp_droprolemember 'largerc', 'loaduser';
```

Exemptions to concurrency limit:

```
CREATE|ALTER|DROP (TABLE|USER|PROCEDURE|VIEW|LOGIN)
CREATE|UPDATE|DROP (STATISTICS|INDEX)
INSERT VALUES
SELECT from system views and DMVs
EXPLAIN
```

Resource class types

Static Resource Classes

Allocate the same amount of memory independent of the current service-level objective (SLO).

Well-suited for fixed data sizes and loading jobs.

Dynamic Resource Classes

Allocate a variable amount of memory depending on the current SLO.

Well-suited for growing or variable datasets.

All users default to the *smallrc* dynamic resource class.

Static resource classes:

```
staticrc10 | staticrc20 | staticrc30 |
staticrc40 | staticrc50 | staticrc60 |
staticrc70 | staticrc80
```

Dynamic resource classes:

```
smallrc | mediumrc | largerc | xlargerc
```

Resource Class	Percentage Memory	Max. Concurrent Queries
smallrc	3%	32
mediumrc	10%	10
largerc	22%	4
xlargerc	70%	1

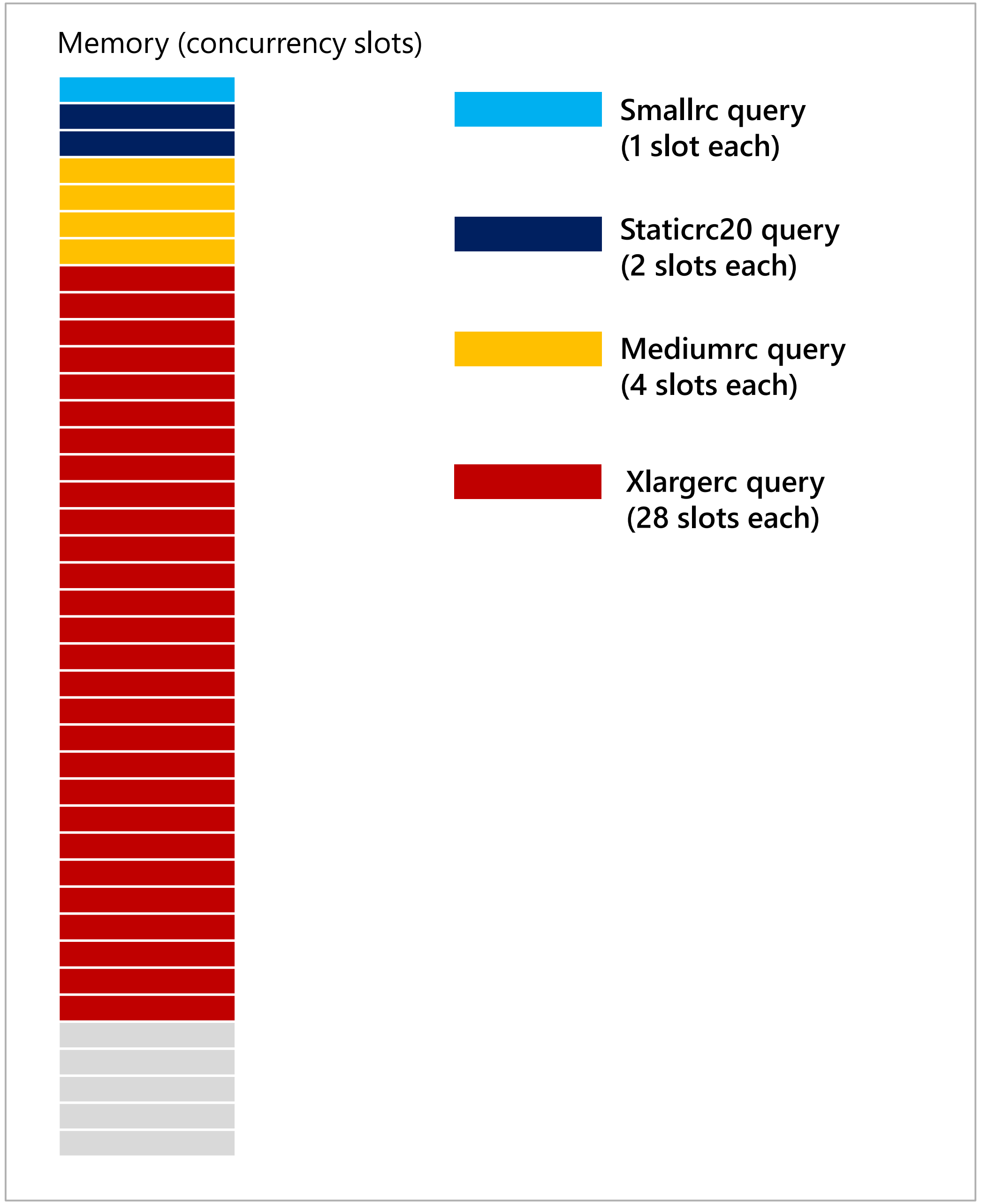
Concurrency slots

Overview

Queries running on a DW compete for access to system resources (CPU, IO, and memory).

To guarantee access to resources, running queries are assigned a chunk of system memory (**a concurrency slot**) for processing the query. The amount given is determined by the resource class of the user executing the query. Higher DW SLOs provide more memory and concurrency slots

@DW1000c: **40 concurrency slots**



Concurrent query limits

Overview

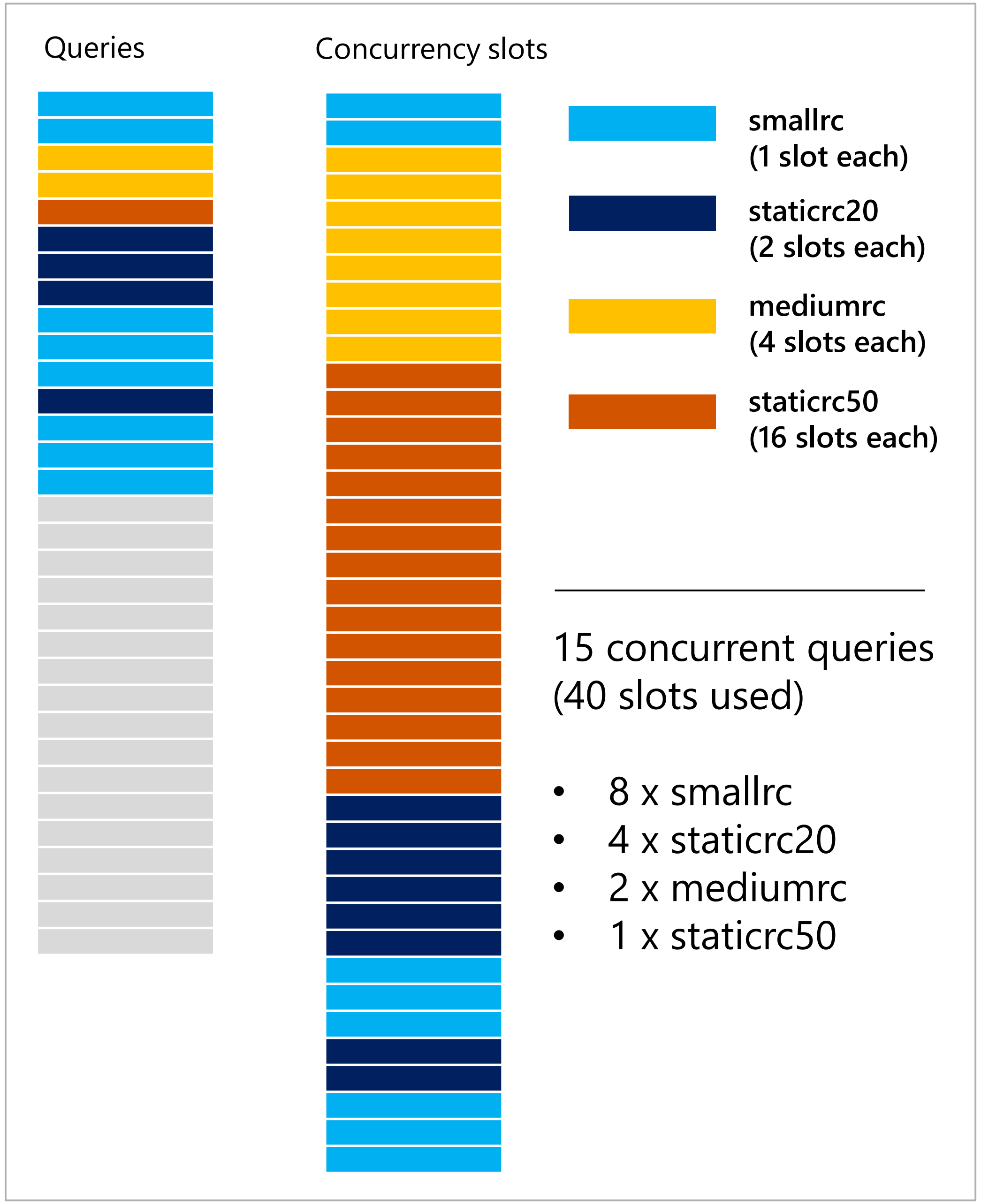
The limit on how many queries can run at the same time is governed by two properties:

- The max. concurrent query count for the DW SLO
- The total available memory (concurrency slots) for the DW SLO

Increase the concurrent query limit by:

- Scaling up to a higher DW SLO (up to 128 concurrent queries)
- Using lower resource classes that use less memory per query

@DW1000c: **32 max concurrent queries, 40 slots**



15 concurrent queries (40 slots used)

- 8 x smallrc
- 4 x staticrc20
- 2 x mediumrc
- 1 x staticrc50

Concurrency limits based on resource classes

Service Level	Max Concurrent Queries	Max Concurrency Slots	Dynamic Resource Classes				Static Resource Classes							
			smallrc	mediumrc	largerc	xlargerc	staticrc10	staticrc20	staticrc30	staticrc40	staticrc50	staticrc60	staticrc70	staticrc80
DW100c	4	4	1	1	1	2	1	2	4	4	4	4	4	4
DW200c	8	8	1	1	1	5	1	2	4	8	8	8	8	8
DW300c	12	12	1	1	2	8	1	2	4	8	8	8	8	8
DW400c	16	16	1	1	3	11	1	2	4	8	16	16	16	16
DW500c	20	20	1	2	4	14	1	2	4	8	16	16	16	16
DW1000c	32	40	1	4	8	28	1	2	4	8	16	32	32	32
DW1500c	32	60	1	6	13	42	1	2	4	8	16	32	32	32
DW2000c	48	80	2	8	17	56	1	2	4	8	16	32	64	64
DW2500c	48	100	3	10	22	70	1	2	4	8	16	32	64	64
DW3000c	64	120	3	12	26	84	1	2	4	8	16	32	64	64
DW5000c	64	200	6	20	44	140	1	2	4	8	16	32	64	128
DW6000c	128	240	7	24	52	168	1	2	4	8	16	32	64	128
DW7500c	128	300	9	30	66	210	1	2	4	8	16	32	64	128
DW10000c	128	400	12	40	88	280	1	2	4	8	16	32	64	128
DW15000c	128	600	18	60	132	420	1	2	4	8	16	32	64	128
DW30000c	128	1200	36	120	264	840	1	2	4	8	16	32	64	128

Workload classification

Overview

Map queries to allocations of resources via pre-determined rules.

Use with workload importance to effectively share resources across different workload types.

If a query request is not matched to a classifier, it is assigned to the default workload group (smallrc resource class).

Benefits

Map queries to both resource management and workload isolation concepts.

Manage groups of users with only a few classifiers.

Monitoring DMVs

`sys.workload_management_workload_classifiers`
`sys.workload_management_workload_classifier_details`
Query DMVs to view details about all active workload classifiers.

```
CREATE WORKLOAD CLASSIFIER classifier_name
WITH
(
   [WORKLOAD_GROUP = '<Resource Class>' ]
   [IMPORTANCE = { LOW                 |
                            BELOW_NORMAL         |
                            NORMAL               |
                            ABOVE_NORMAL         |
                            HIGH
                      }
   ]
   [MEMBERNAME = ‘security_account’]
)
  WORKLOAD_GROUP: maps to an existing resource class
  IMPORTANCE: specifies relative importance of
              request
  MEMBERNAME: database user, role, AAD login or AAD
              group
```

Workload importance

Overview

Queries past the concurrency limit enter a FiFo queue

By default, queries are released from the queue on a first-in, first-out basis as resources become available

Workload importance allows higher priority queries to receive resources immediately regardless of queue

Example Video

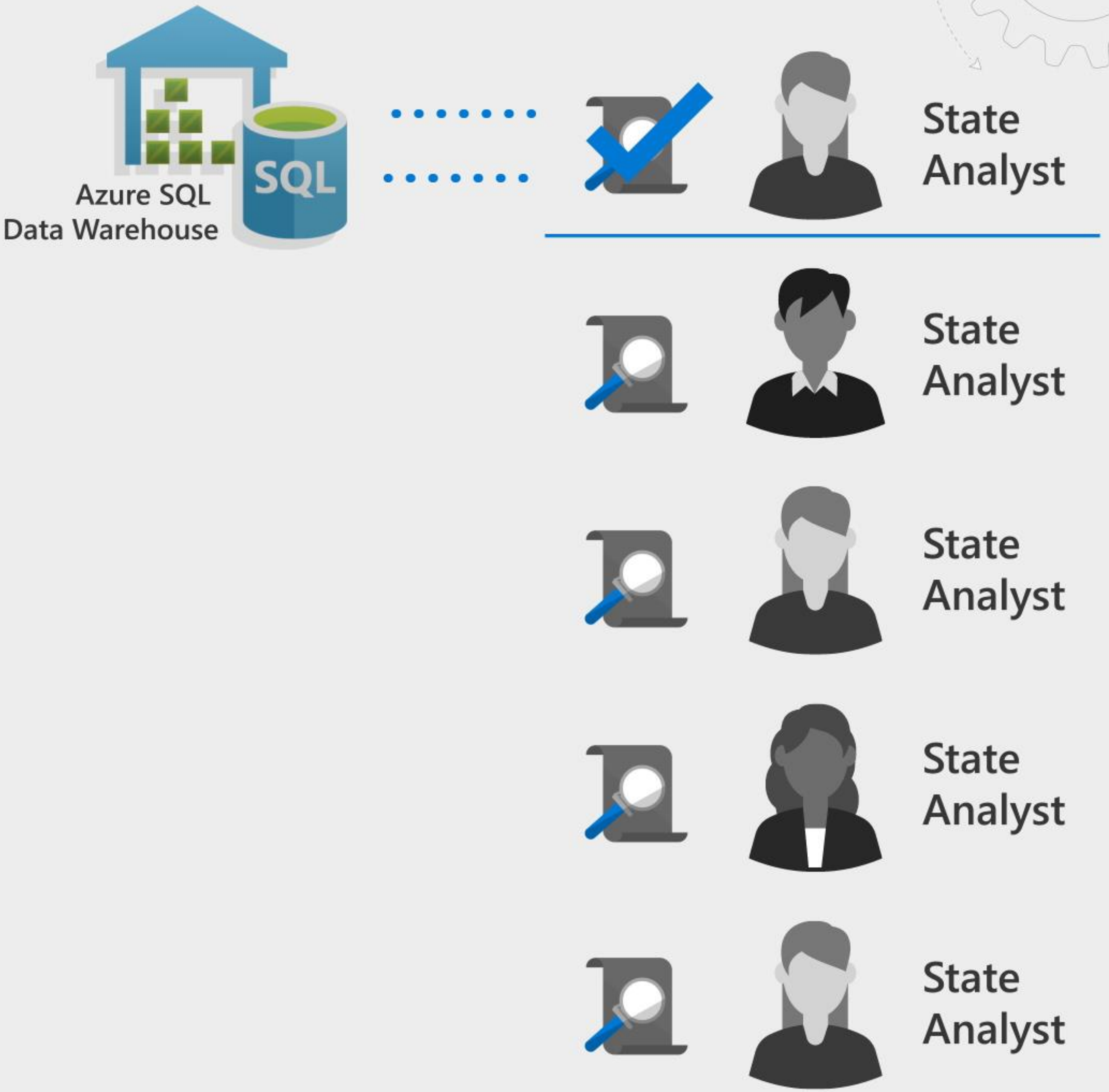
State analysts have normal importance.

National analyst is assigned high importance.

State analyst queries execute in order of arrival

When the national analyst’s query arrives, it jumps to the top of the queue

```
CREATE WORKLOAD CLASSIFIER National_Analyst
WITH
(
   [WORKLOAD_GROUP = ‘smallrc’]
   [IMPORTANCE = HIGH]
   [MEMBERNAME = ‘National_Analyst_Login’]
```


SQL
Azure SQL Data Warehouse
State Analyst
State Analyst
State Analyst
State Analyst
State Analyst