Technical Brief

Qlik® Sense Performance Benchmark

This technical brief outlines performance benchmarks for Qlik Sense and is based on a testing methodology called the Qlik Capacity Benchmark. This series of tests is conducted by varying data volumes, users, and applications in a given environment and recording the results. This exhaustive set of permutations yields a matrix of CPU utilization, RAM utilization, and response times. This approach is different than many other scalability tests. Not only are metrics reported when a server is saturated, but also when the server is only partially utilized. This methodology provides transparency to the testing process and resulting metrics. While Qlik applications do vary in size and complexity, the methodology provides a comprehensive set of data with which customers can judge scalability and plan for deployments.

These performance benchmarks are applicable to Qlik Sense Enterprise and Qlik Analytic Platform.

For more information about Qlik Sense, please reference the **Qlik Sense architectural overview whitepaper**, and the **Qlik Sense scalability datasheet.**

Quick Facts

This benchmark demonstrates capacity and performance with variations of concurrent users, data volumes, and applications complexities listed below.

Scenario	Software	Engine Cores	Concurrent Users	Data Volumes	Application Complexity
1	Qlik Sense 3.2	20	10 - 500	10 – 500M rows	moderate, complex
2	Qlik Sense 3.2	8 vCPU	10 - 200	10 – 200M rows	moderate, complex
3	Qlik Sense 3.2	4 vCPU	10 - 125	10 – 100M rows	moderate, complex
4	Qlik Sense 3.2	4 vCPU	1000	1M rows	simple

Additionally, the following capabilities were explored.

Scenario	Description
5	Associative Search
6	Result Caching

Hardware

The following environment was used to perform the test:

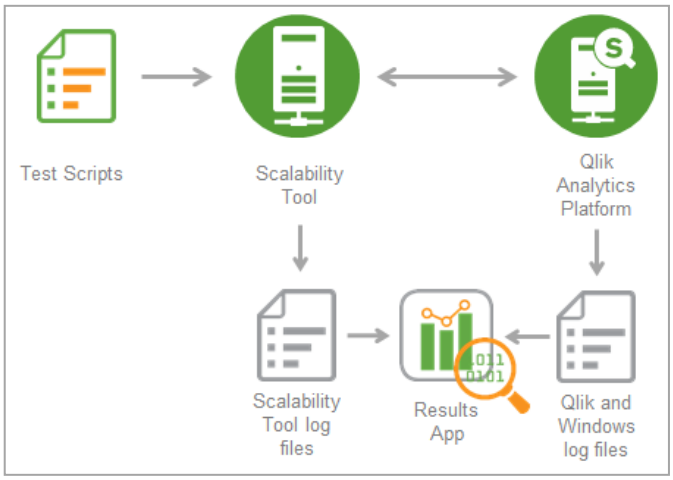
Scenario	Software	Hardware	Processor	RAM
1	Qlik Sense 3.2	HP DL380 Gen9	E5-2687wv3	384 GB
2 - 4	Qlik Sense 3.2	AWS r3.2xlarge	8 vCPU / E5-2670 v2 chipset	61 GB
All	Qlik Sense Scalability Tool Load Client	AWS r3.2xlarge	8 vCPU / E5-2670 v2 chipset	61 GB

Notes about the test configuration:

- HP DL380 Gen9 has hyper-threading enabled. It is a 20 physical core server with 40 virtual cores.
- In Scenario #3, Engine activity is restricted to four cores. As a best practice, an eight core VM was used to allow for other processes (e.g., Proxy, Repository) to have capacity with which to operate.
- These performance benchmarks are applicable to Qlik Sense Enterprise and Qlik Analytic Platform.

Qlik Sense Scalability Tool

Qlik Sense can be load tested with a freely available load testing tool called Qlik Sense Scalability Tool. It is found here: https://community.qlik.com/docs/DOC-8878. This load testing client can simulate virtual users against customer Qlik Sense applications.



Using the Qlik Sense Scalability Tool, test scripts simulate virtual users and are executed against Qlik Sense applications. Upon completion, performance metrics from Qlik Sense, Windows, and the Scalability Tool are collected into a Qlik application for analysis. All virtual users were simulated to be highly interactive with the application. In all scenarios, virtual users interacted with charts and filter panes, navigated among tabs, and performed actions within applications. This provided a realistic view of how Qlik Sense handles a given user load. Virtual users were simulated with 20 to 30-second think times and made randomized selections throughout the tests, rather than the same selection, to minimize caching that might underreport utilization averages. Each test ran for one hour and reached full load in 20 minutes.

Control Variables

The control variables varied throughout a series of one hour tests from the below:

- **Applications**: moderate, complex
- **Concurrent Users**: 10, 25, 50, 75, 100, 200, 500
- **Data Volumes (Millions)**: 10, 50, 100, 200, 500

Metrics

The metrics captured during each one hour tests are below:

- **Average Engine CPU Utilization**: 0-100%
- **Max RAM Utilization**: 0 GB – Max GB of Server
- **Average User Response Time**: 0 sec – 5(+) sec

Test #	Application	Concurrent Users	Data Volume
1	moderate	10	10
2	moderate	50	10
…	…	…	…
6	moderate	10	50
…	…	…	…
12	complex	10	10
…	…	…	…
56	complex	50	500

Results

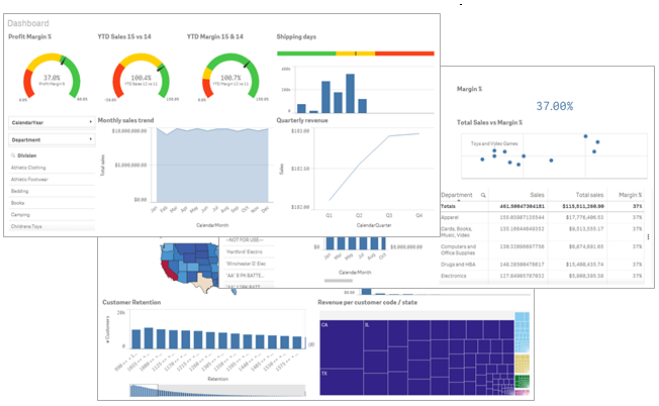
Test #	CPU %	RAM (GB)	Response Time
1	4	4	0.5
2	8	5	0.5
…	…	…	…
6	12	12	0.5
…	…	…	…
12	6	6	0.8
…	…	…	…
56	65	200	1

Example Capacity Benchmark Matrix

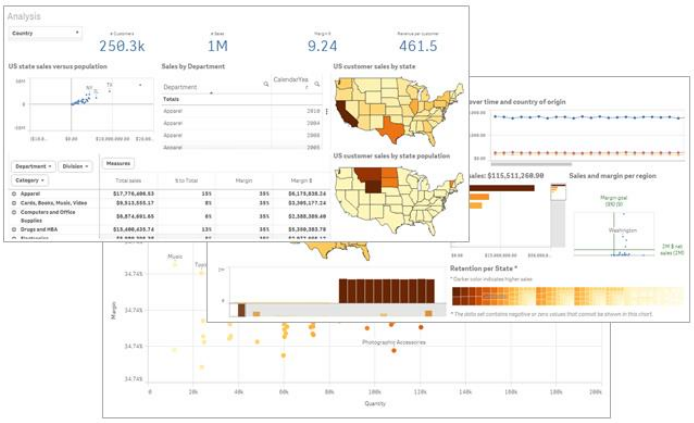
Applications

The complexity of an application influences how many concurrent users and how much underlying data it can support. Qlik Sense applications range from simple lookups of information to complex visualizations, calculations, and use cases. The Qlik Capacity Benchmark tests account for this variation by testing applications with different presentation layers, calculations, and test scripts.
The applications used in the tests are benchmark apps designed to reflect common scenarios that we see across our customers.

Sales Dashboard (Moderate)
The Sales Dashboard application shows data in aggregate via many graphical objects. In addition to gauges and trends, it allows for more complex analysis including mapping, scatter charts, and set analysis. The client load script simulates a use case where users research data at an aggregate level, drill through many contexts (customer, profitability), and interact with charts.



Sales Analysis (Complex)
The Sales Analysis application analytics including many graphical objects and calculations per sheet, dense scatter plots, and some detail data throughout the application. The client load script simulated a use case where users perform complex analytics, including geospatial analysis, set analysis, and product ranking.



Retail Website (Simple)
The Retail Website application shows a simple scenario where the user researches product location, availability, and pricing on a limited dataset of 1 million rows. This application is not explored in the Capacity Benchmark sections below but is found in a later section of this document.

Overall Score

The results from the Capacity Benchmark are categorized according to the thresholds defined in the table to the right. The primary metrics of average engine CPU utilization, maximum RAM utilization, and average response times are scored in this way to provide visual feedback about the overall performance of the server in each scenario. The tests are unbounded and are not intended to show maximum performance. The scores also give an indication of the overall remaining capacity of the server in each configuration. Finally, the scores are rolled into an overall score for the server.

Score	CPU	RAM	Response Time
🟢	< 60 %	< 70%	< 1 s
🟡	< 80 %	< 90 %	< 3 s
🔴	> 80%	> 90 %	> 3 s

In the results below, for example, a green mark indicates that the test completed with less than 60% average CPU utilization, 70% RAM utilization, and less than one second response time. A yellow mark indicates one or more metrics entered the yellow range, and a red marking indicates one or more metrics entered the red range.

In addition, this document doesn't comment on the total *addressable* data per server. With application architectures, such as On-Demand-App-Generation (ODAG) and Direct Discovery, the total addressable amount of source data could be much larger. For more information please see the Qlik and Big Data whitepaper and also visit Qlik Community.

Note there is nothing inherently wrong with a server running more than 60% CPU or 70% RAM utilization; we do this to show a realistic viewpoint of the remaining server capacity under a given load from a sizing and capacity planning standpoint. As shown below, tests with 'yellow' CPU or RAM utilization still yield acceptable response times, but may be limited in additional capacity. 'Red' indicates system saturation.

Scenario #1 - Capacity Benchmark Summary – 20 physical cores

HP DL380 Gen9, 20 physical cores, E5-2687w v3, 384 GB RAM, Hyper-threading enabled

Qlik Sense was tested with 50 one hour performance tests split equally between moderate and complex applications. The overall scores are displayed to the right where each point represents an hour-long test.

Results

Qlik Sense was able to reach 750 concurrent users (7500+ total users) on a 50 millon row data set and 10-50 concurrent users (100-500+ total users) on a 500 millon row data set.

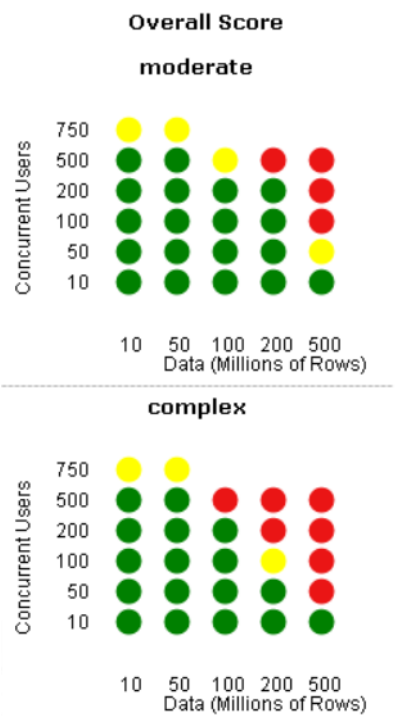
Data High Water Mark

Application Complexity	Data Volume (Millions of Rows)	Concurrent Users	Avg Response Time (Seconds)	Average CPU %	Max RAM (GB)
moderate	500	50	1.1	36%	178
complex	500	10	1.0	26%	103

User High Water Mark

Application Complexity	Data Volume (Millions of Rows)	Concurrent Users	Avg Response Time (Seconds)	Average CPU %	Max RAM (GB)
moderate	50	750	1.9	29%	102
complex	50	750	0.8	51%	138

Overall score – 20 cores



Scenario #2 - Capacity Benchmark Summary – 8 cores

AWS r3.2xlarge, 8 vCPU, E5-2670 v2, 61 GB RAM, Hyper-threading enabled

Qlik Sense was tested with 42 one hour performance tests split equally between moderate and complex applications. The overall scores are displayed to the right where each point represents an hour-long test.

Results

Qlik Sense was able to reach 125-200 concurrent users (1250-2000+ total users) on a 10 millon row data set and 25-50 concurrent users (250-500 total users) on a 200 millon row data set.

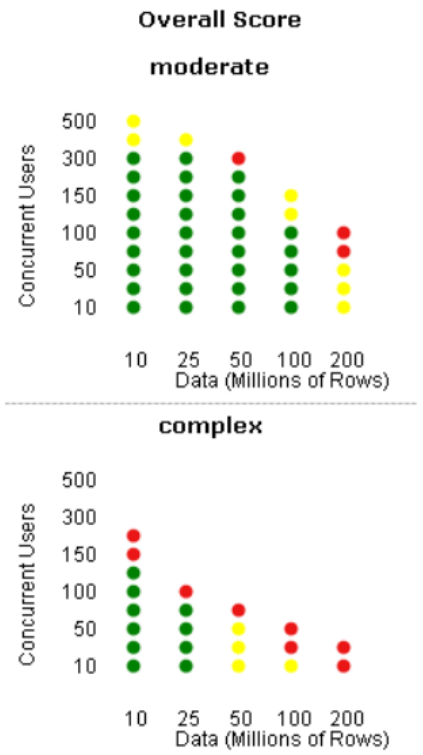
User High Water Mark

Application Complexity	Data Volume (Millions of Rows)	Concurrent Users	Avg Response Time (Seconds)	Average CPU %	Max RAM (GB)
moderate	10	125	0.3	11%	15
complex	10	200	0.5	47%	28

Data High Water Mark

Application Complexity	Data Volume (Millions of Rows)	Concurrent Users	Avg Response Time (Seconds)	Average CPU %	Max RAM (GB)
moderate	200	50	3.0	49%	43
complex	200	25	18.7	88%	42

Overall score – 8 cores



Scenario #3 - Capacity Benchmark Summary – 4 cores

AWS r3.2xlarge, 8 vCPU, E5-2670 v2, 61 GB RAM, Hyper-threading enabled

In this scenario, Engine activity is restricted to 4 cores. As a best practice, an 8 core VM was used to allow for other processes (e.g., Proxy, Repository) to have capacity with which to operate. Qlik Sense was tested with 34 one hour performance tests split equally between moderate and complex applications. The overall scores are displayed to the right where each point represents an hour-long test.

Result

Qlik Sense was able to reach 125-400 concurrent users (1250-4000+ total users) on a 10 millon row data set and 25-150 concurrent users (250-1500+ total users) on a 100 millon row data set.

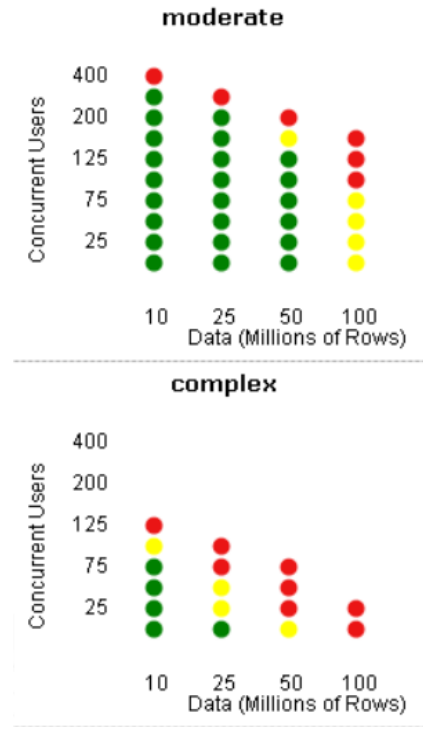
User High Water Mark

Application Complexity	Data Volume (Millions of Rows)	Concurrent Users	Avg Response Time (Second...	Average CPU %	Max RAM (GB)
moderate	10	400	13.8	29%	28
complex	10	125	4.6	44%	26

Data High Water Mark

Application Complexity	Data Volume (Millions of Rows)	Concurrent Users	Avg Response Time (Seconds)	Average CPU %	Max RAM (GB)
moderate	100	150	15.0	43%	39
complex	100	25	15.2	47%	30

Overall score – 4 cores



Scenario #4 - Retail Website Application – 4 cores

AWS r3.2xlarge, 8 vCPU, E5-2670 v2, 61 GB RAM, Hyper-threading enabled

In this scenario, Engine activity is restricted to 4 cores. As a best practice, an 8 core VM was used to allow for other processes (e.g., Proxy, Repository) to have capacity with which to operate.

The Retail Website application shows a simple scenario where the user researches product location, availability, and pricing on a limited dataset of 1 million rows. This scenario is intended to illustrate the scale at which Qlik operates scenarios where the data set is very small.

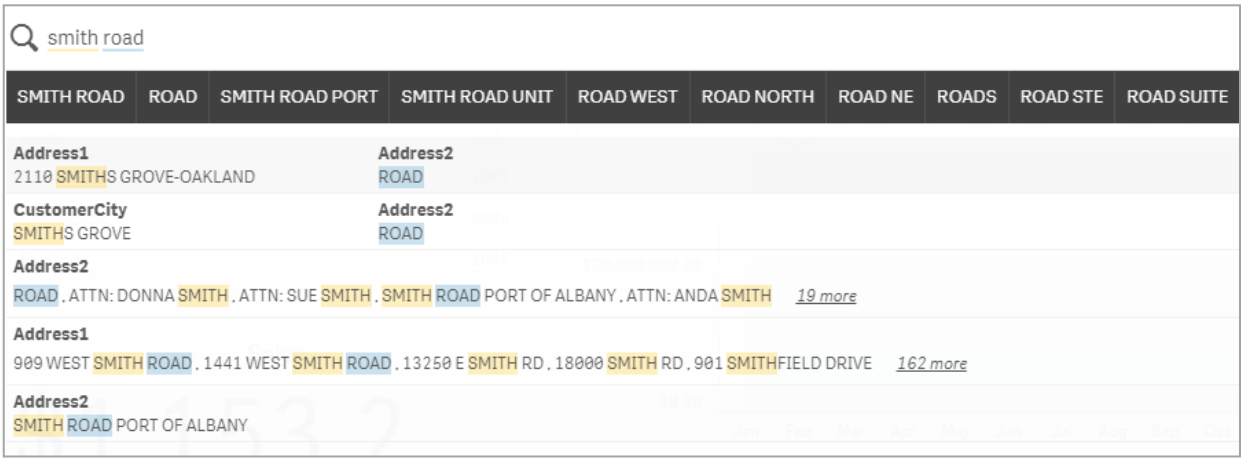
Test Input
1 million rows of data
1000 concurrent users

Metric	Value
Engine CPU %	23%
Average Response time	1.0 seconds

Scenario #5 – Associative Search

HP DL380 Gen9, 20 physical cores, E5-2687w v3, 384 GB RAM, Hyper-threading enabled

Associative Search is a Qlik Sense capability that allows users to search with many search terms across many tables and fields in their application simultaneously. This enables users to quickly search for and find relevant data quickly. In the screenshot below, the values “smith” and “road” are searched and Qlik Sense displays results from several tables and fields simultaneously.



While it is possible to search across all fields in a data model, it often makes sense to restrict the search fields to those that are relevant to the use case. In this scenario, 20 search fields were chosen across six tables (five dimension tables and one fact table), with 50,611,089 distinct values in total.

Test Input
50 million rows of data
Search 6 tables, 20 fields

Result Metric	Value
Average Engine CPU %	28%
Average Response time	1.2 seconds
Throughput	450 distinct searches / minute

Consider the complexity of SQL required to implement this using traditional SQL based analytics. Qlik's Associative Indexing (QIX) Engine indexes all the data to make this possible and delivers an end user experience where users can answer more questions without IT involvement.

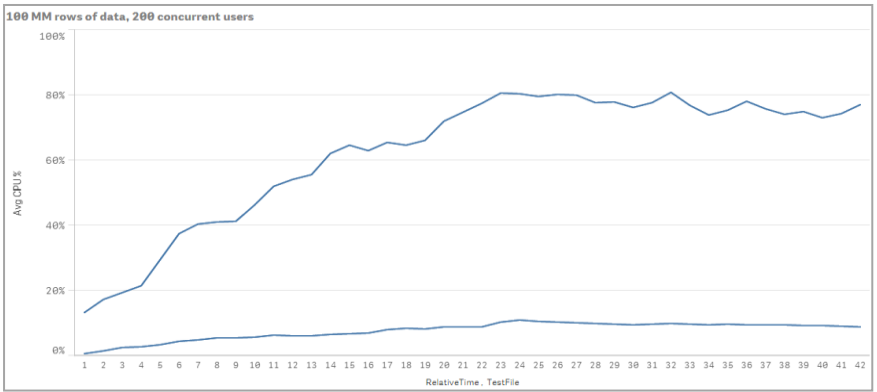
Scenario #6 – Result Caching

In the scenarios above, virtual user scenarios were designed to minimize the use of Qlik's result caching ability by completely randomizing the selections virtual users made. This is to exercise the QIX engine to the greatest degree during each scenario. This section now looks at the effects of result caching. As a user makes a selection, the results from the selection state (i.e., associative; the green, white, and grey indication of relationships among data) and chart calculations themselves are cached. Any identical subsequent selections result in selection state and chart calculation retrieval from the result cache. This results in improved response time and reduced CPU utilization that further extends Qlik's performance and capacity. The metrics below provide a clear indication of the efficiency of Qlik's result cache.

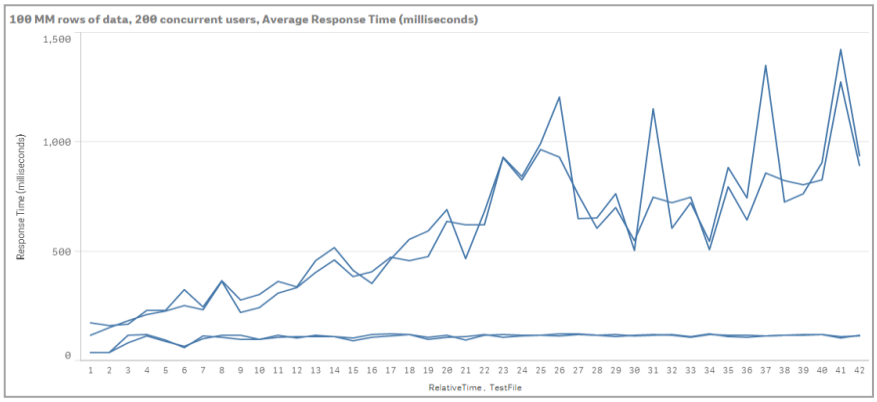
Test Input
100 million rows of data
200 concurrent users

Result Metric	Non-Cached	Cached
Average Engine CPU %	~80%	< 10%
Average Response time	~1 second	~0.3 second

CPU utilization is reduced from 80% to 10% as the result cache is leveraged.



Average response times drop from ~1 second to ~0.3 seconds as the result cache is leveraged. ~0.3 seconds represents the time it takes to fetch a result from the cache regardless of how long it took the original calculation to complete.



The cache hit rate percentage is completely dependent upon the application and the use case, and so it may be that the cache hit rate percentage for another application and use case is less. Still, this provides a clear example of the effectiveness of Qlik's result caching ability and to the degree an application and use case produces repeat calculations it will have a corresponding reduction in CPU utilization and response times.

Conclusion

The Qlik Capacity Benchmark tests are different from many other scalability tests. Not only is a clear indication given around the data volumes and concurrent users that Qlik products can handle when a server is taken to saturation, but these tests also show metrics when a server is not saturated, as well. These fundamental and critical metrics of CPU, RAM, and response times provide a complete and transparent view of the performance of Qlik Sense.

150 N. Radnor Chester Road
Suite E120
Radnor, PA 19087
Phone: +1 (888) 828-9768
Fax: +1 (610) 975-5987

qlik.com