

Capacity Requirements for Web Analytics Shared Service in SharePoint Server 2010

This document is provided “as-is”. Information and views expressed in this document, including URL and other Internet Web site references, may change without notice. You bear the risk of using it.

Some examples depicted herein are provided for illustration only and are fictitious. No real association or connection is intended or should be inferred.

This document does not provide you with any legal rights to any intellectual property in any Microsoft product. You may copy and use this document for your internal, reference purposes.

© 2010 Microsoft Corporation. All rights reserved.

**Capacity Requirements for Web Analytics Shared Service in SharePoint Server 2010**

|  |  |
| --- | --- |
| Authors | Tushar Garg |
| Testers | Shuang Yang |

Microsoft Corporation

April 2010

**Applies to: Microsoft SharePoint Server 2010**

**Summary**: This white paper provides guidance on capacity management for Microsoft® SharePoint® Server 2010 Web Analytics shared service.

Following aspects of capacity planning are described below for the Web Analytics shared service:

* Description of the architecture and topology.
* Capacity planning guidelines based on the key factors such as total expected traffic and number of SharePoint components.
* Description of the other factors that affect the performance and capacity requirements.

Contents

[**Prerequisite information** 4](#_Toc262816028)

[**Introduction** 5](#_Toc262816029)

[**Overview** 5](#_Toc262816030)

[**Architectural overview** 5](#_Toc262816031)

[**Specifications** 7](#_Toc262816032)

[**Hardware** 7](#_Toc262816033)

[**Topology** 9](#_Toc262816034)

[**Capacity requirements** 10](#_Toc262816035)

[**Testing methodology** 10](#_Toc262816036)

[**Dataset description** 11](#_Toc262816037)

[**Application servers** 12](#_Toc262816038)

[**SQL Server-based computers** 13](#_Toc262816039)

[**Other factors** 14](#_Toc262816040)

[**Remaining issues:** 15](#_Toc262816041)

**Prerequisite information**

Before reading this document, it is important that you understand the key concepts behind Microsoft® SharePoint Server 2010 capacity management. The following documentation will help you learn about the recommended approach to capacity management and provide context for helping you understand how to make effective use of the information in this document, as well as define the terms used throughout this document.

For more conceptual information about performance and capacity that that you might find valuable in understanding the context of the data in this technical case study, see the following documents:

* [Capacity Planning and Sizing for Microsoft SharePoint 2010 Products and Technologies](http://technet.microsoft.com/en-us/library/cc262971(Office.14).aspx)
* [SharePoint Server 2010 Software Boundaries](http://technet.microsoft.com/en-us/library/cc261700(Office.14).aspx)

**Introduction**

**Overview**

Web Analytics service as part of Microsoft® SharePoint® Server 2010, is set of features to help you collect, report, and analyze the usage and effectiveness of your SharePoint Server 2010 deployment. There are three categories of the SharePoint Web Analytics reports: Traffic, Search, and Inventory. The reports are aggregated for various SharePoint entities like Site, Site Collection, and Web Application for each farm. The architectural overview of the Web Analytics service in a SharePoint deployment is described in Figure1 in Architectural Overview section below.

The Web Analytics shared service requires resources primarily at the application server and SQL Server level. This document does not cover the Web Server layer capacity planning, because the Web Analytic Service’s capacity requirements are minimal at this level.

This document covers the capacity requirements for the number of application servers and SQL Server-based computers, based on the following criteria:

1. Total expected site traffic (clicks, search queries , ratings).
2. Number of SharePoint components (Site, Site Collection, and Web Application) for each farm.

Other less significant factors which can affect the capacity requirements are summarized in the Other Factors.

**Architectural overview**

Figure 1, later in this white paper, describes the flow of the site usage data from the Web Browser to the Analytics databases and back onto the Web Browser as reports. The usage data is logged to the usage files on the Web Serves. The Usage timer job calls the Logging Web Service to submit the raw data from the usage files. The Logging Web Service writes it to the staging database, where the raw data is stored for seven days (not configurable). The Web Analytics components Log Batcher and User Behavior Analyzer clean and process the raw data on the staging database. The Report Consolidator runs once every 24 hours and aggregates the raw data from the staging database on various dimensions and writes it to the reporting database. The aggregated data is stored in the reporting database for a default period of 25 months (configurable).

The performance of the Logging Web Service’ primarily depends on the number of application servers. Scaling out is available for the application servers. The performance of the Log Batcher and User Behavior Analyzer depends primarily on the Analytics staging database. The read and write by all these components makes the Analytics staging database a bottle neck. Scaling out is available for the staging database. The performance of the Report Consolidator also primarily depends on the Reporting database. However, scaling out of Reporting database is not supported. It should be noted that the same server with SQL Server can be used to deploy both the analytics staging database and the reporting database along with the other SharePoint databases.



**Figure1. SharePoint Server 2010 Web Analytics Architectural Overview**

**Specifications**

This section provides detailed information about the hardware, software, topology, and configuration of the case-study environment.

**Hardware**

|  |  |
| --- | --- |
| Note |  |
| This environment is scaled to accommodate pre-release builds of SharePoint Server 2010 and other products. Hence, the deployed hardware has greater capacity than necessary to serve the demand typically experienced by this environment. This hardware is described only to provide additional context for this environment and serve as a starting point for similar environments.  It is important to conduct your own capacity management based on your planned workload and usage characteristics. For more information about the capacity management process, see [Performance and capacity management (SharePoint Server 2010)](http://technet.microsoft.com/en-us/library/cc262971(Office.14).aspx). | |

**Web servers**

This document does not cover the Web Server layer capacity planning, because the Web Analytic service’s capacity requirements are minimal at this level.

**Application servers**

The following table describes the configuration of each of the application servers. Based on the site traffic and number of SharePoint components the users will need one or more application servers.

|  |  |
| --- | --- |
| Application server |  |
| Processor(s) | 2 quad core @ 2.33 GHz |
| RAM | 8 GB |
| Operating system | Windows Server® 2008, 64 bit |
| Size of the SharePoint drive | 300 GB |
| Number of NICs | 1 |
| NIC Speed | 1 Gigabit |
| Authentication | NTLM |
| Load balancer type | SharePoint Load Balancer |
| Software version | SharePoint Server 2010 (pre-release version) |
| Services running locally | Central Administration  Microsoft SharePoint Foundation Incoming E-mail  Microsoft SharePoint Foundation Web Application  Microsoft SharePoint Foundation Workflow Timer Service  Search Query and Site Settings Service  SharePoint Server Search  Web Analytics Data Processing Service  Web Analytics Web Service |
|  |  |

**Database servers**

We need instances of SQL Server both for the staging and reporting databases.

|  |  |
| --- | --- |
| Database Server | DB1-2 |
| Processor(s) | 4 quad core @ 2.4 GHz |
| RAM | 32 GB |
| Operating system | Windows Server 2008, 64-bit |
| Disk Size | 3 terabytes |
| Number of NICs | 1 |
| NIC Speed | 1 Gigabit |
| Authentication | NTLM |
| Software version | SQL Server® 2008 |

**Topology**



**Capacity requirements**

**Testing methodology**

This document presents the capacity requirements in terms of the total amount of site traffic (clicks, search queries, ratings) per day that can be supported by different numbers of application servers and SQL Server-based computers. The numbers presented currently are for a midsize SharePoint deployment with about 30,000 SharePoint entities. The Web Analytics shared service aggregates the data for each day and thus the data volume is presented corresponding to the total number of records (clicks, searches, ratings) each day that the SharePoint farm is expected to receive.

The numbers are presented in two colors:

1. Green – Indicating the safe limit for the site traffic that can be processed for the corresponding number of application servers and SQL Server-based computer.
2. Yellow – Indicating the expected limit for the site traffic that can be processed for the corresponding number of application servers and SQL Server-based computers.

The green and yellow numbers are an estimates based on the following key factors -

1. Total Site Traffic (Page view clicks, Search queries and clicks, ratings).
2. Number of SharePoint entities (Site, Site Collection, and Web Application) for each farm.

The estimates also depend on other properties of the data and the data retention period in the reporting database. For the purpose of the testing, we have maintained the other properties of the data as constant as described in the following [Dataset Description](#_Data_Set_Description) section.

Also, although in the SharePoint deployment, particularly in smaller deployments, it is quite possible to share the application servers and SQL Server-based computers with other SharePoint services and databases. This document provides the capacity of the application servers and the SQL Server-based computers, based on the test environment where Web Analytics shared service is the only major service running on the servers. The actual performance results in the environments with other shared services actively and simultaneously running could vary.

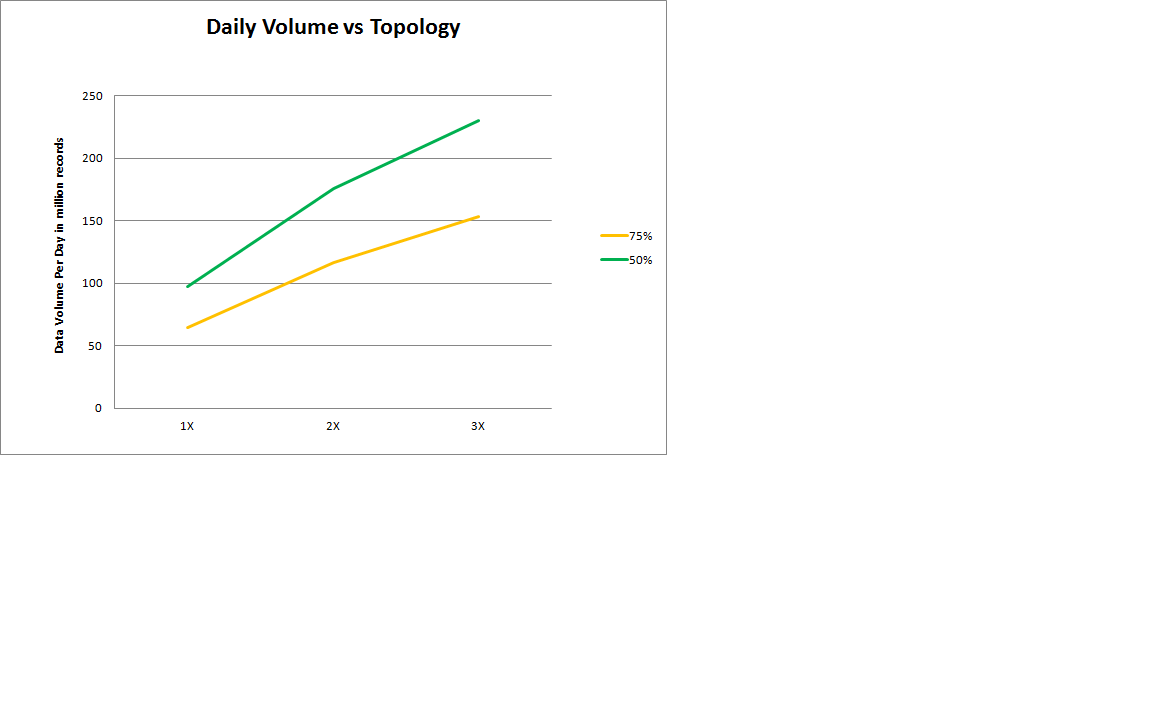
To decide on the capacity requirements, the expected daily site traffic and the number of components in the SharePoint deployment should be estimated. Thereafter, the number of application servers and SQL Server-based computers should be estimated independently, based on the following Figure 3 and Figure 4, respectively.

**Dataset description**

The dataset that we chose for the experiment is the midsize dataset with approximately 30,000 SharePoint components. The other characteristics of the data are also mentioned in the following table and were kept constant across the experiment.

|  |  |  |
| --- | --- | --- |
| Dataset characteristics | Value | |
| Number of SharePoint components | | 28,967 |
| Number of unique users | | 117,000 |
| Number of unique queries | | 68,000 |
| Number of unique assets | | 500,000 |
| Data size in the reporting database | | 200 GB |

The total site traffic (clicks, searches, ratings) was increased as part of the experiment, to establish the number of records that can be supported for the corresponding topology.



**Application servers**

The following diagram shows the daily site traffic that can be supported, corresponding to 1, 2 or 3 Application Servers, respectively. The site traffic is represented in millions of records (each click, search, or rating constitutes a record) each day. The yellow line represents the expected number of records for the corresponding topology, while the green line represents the safe assumption for the number of records.

**Figure3. Daily Site Traffic vs. the Application Servers Topology**

The application servers are not very CPU intensive or memory intensive. Thus the CPU and the memory usage are not summarized for this section.

**SQL Server-based computers**

The diagram below shows the daily site traffic that can be supported corresponding to 1S+R (one instance of SQL Server for both staging and reporting databases), 1S1R (two instances of SQL Server, one staging database and one reporting database) and 2S1R (three instances of SQL Server, two staging databases and one reporting database). The site traffic is represented in millions of records (each click, search, or rating constitutes a record) each day. The yellow line represents the expected number of records for the corresponding topology, while the green line represents the safe assumption for the number of records.

**Figure3. Daily Site Traffic vs. SQL Server Topology**

The following table summarizes the CPU and memory usage of the various components on the instances of SQL Server that are hosting the staging database and the reporting database.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Configuration** | **1S+R** | **1S1R** | **1S1R** | **2S1R** | **2S1R** |
|  | **Staging + Reporting** | **Staging** | **Reporting** | **Staging** | **Reporting** |
| **Total sum of percentage of processor time for 8 processor computer** | 19 | 192 | 5.78 | 100 | 13.4 |
| **SQL Server Buffer Hit Ratio** | 99 | 100 | 100 | 100 | 100 |
| **%Disk time** | 7,142 | 535 | 5.28 | 59.3 | 98.2 |
| **Disk Queue Length** | 357 | 28.6 | 0.26 | 2.97 | 4.91 |

**Other factors**

There are numerous other factors that can affect the performance of various analytics components and can affect the capacity planning. These factors primarily affect the performance of the Report Extractor component because they can affect the size of the data aggregated each day. The total size of the data in the reporting database also affects the performance of the Reporting Extractor, although it is not very significant because the data is partitioned daily. Some of these factors are:

1. Number of unique queries each day
2. Number of unique users each day
3. Total number of unique assets clicked each day
4. Existing data size in the reporting warehouse, based on the data retention in Warehouse

The overall effect of such factors is less significant than the total data volume and the number of site entities. However, it is important to conduct your own capacity management based on your planned workload and usage characteristics. For more information about the capacity management process, see [Performance and capacity management (SharePoint Server 2010)](http://technet.microsoft.com/en-us/library/cc262971(Office.14).aspx)

**Remaining issues:**

There are current known issues which significantly affect the current performance of Web Analytics Service for deployments with a large site hierarchy (approximately 100,000 SharePoint components and more). This document will be updated with the capacity requirements for large site hierarchies after the issue is fixed.