ESI.Lab2 - Estudio de rendimiento de un servidor Web

Introducción

La calidad del servicio de un sistema informático (*IT system*) se puede determinar en base a los índices de disponibilidad, fiabilidad y rendimiento obtenidos a partir de los datos de actividad. En particular, en el caso de los servicios web, el análisis de la actividad y la caracterización de la carga del sistema tiene una gran relevancia en la evaluación, el ajuste de los niveles de servicio, y en la estimación de la capacidad necesaria para garantizar los niveles de servicio en escenarios de aumento de la carga. En el caso de los servicios web la demanda de servicio tiene un fuerte componente aleatorio ya que ni el número de usuarios, ni la forma en la que solicitan servicios, ni los patrones temporales de utilización, pueden ser completamente determinados o conocidos a priori.

Teniendo en cuenta las características de la demanda a los servicios web, la evaluación del rendimiento de un servidor web ha de tomar en consideración el patrón temporal al que se ajusta la demanda de servicio, los parámetros que caracterizan la intensidad de la carga de dicha demanda, el análisis de las respuestas del servidor y su comportamiento (utilización de recursos) en función de la intensidad de la demanda. Los patrones de trafico no solamente tendrán consecuencias en la caracterización de la carga y la determinación de la capacidad actual del servidor, sino que serán la base para establecer las previsiones de utilización y demanda de los recursos del sistema. En la caracterización de las respuestas del servidor se determinan los tamaños y tiempos de las respuestas, productividades, latencias y utilizaciones del sistema y de sus componentes.

Hay que señalar que en el rendimiento de los servidores web, además de la frecuencia y tipo de las peticiones al servidor, intervienen entre otros los siguientes factores:

- clientes, servidores, proxies, networks y protocolos.
- File Caching en los navegadores web clientes para reducir las cargas.
- Web proxies para reducir los tiempo de respuesta y el tráfico de red.
- Caché de red para los ficheros para reducir el tráfico de internet.

Disponer de información completa acerca de las características de estos elementos, además de información de la demanda en el sistema de las transacciones http analizadas, permitirá elaborar modelos mucho más precisos acerca de los efectos que las peticiones de usuario tendrían en la utilización de recursos y rendimiento del sistema para una carga de trabajo dada.

Práctica LRdto.CWeb - Evaluación del rendimiento de un servidor web

La práctica a desarrollar consistirá en el análisis de la actividad y la caracterización de la carga de un servidor web. El análisis y caracterización de la carga se realizará a partir de la información recogida en el fichero de registro de la actividad (*log file*) de un servidor web. Dada la naturaleza de la información disponible la caracterización de la carga se realizará en base a la popularidad de los documentos accedidos, del tamaño de los ficheros, del tiempo consumido en servir la petición y de los patrones de peticiones de usuario. Las conclusiones y el alcance del estudio estarán muy limitadas al no disponer de información precisa acerca de la demanda de recursos (cpu, disco, red, ...) que cada petición genera en el servidor.

Objetivos

- Utilización de técnicas de evaluación de la actividad de un servidor web.
- Puesta en práctica de las técnicas y métodos de caracterización de la carga vistas en la asignatura.
- Familiarizarse con la elaboración del modelo de carga de un servidor web.

Desarrollo de la práctica

Para el desarrollo de la práctica se recomienda la metodología descrita en el libro de Molero et al. [1, cap. 8] para las fases de construcción del modelo de carga.

Sistema en estudio.

Se realizará el análisis y caracterización de la carga para la actividad recogida en el servidor web de la Escuela de Ingeniería Informática de Valladolid (https://www.inf.uva.es/) entre el 18 de Octubre de 2017 y el 3 de Diciembre de 2017.

Información disponible

En un servidor web cada transacción de protocolo http, se complete o no, se recoge en alguno de los ficheros de registro de actividad (*web server log files* o sencillamente *log files*) del servidor. Estos ficheros son almacenados en texto plano (ASCII), independientes de plataforma, y con los campos separados por tabuladores o espacios. En un servidor web se pueden encontrar cuatro tipos de ficheros de registro de la actividad:

- Transfer (access) log
- Error Log
- Referer Log
- Agent Log.

Los datos a analizar en esta práctica son los que se recogen en el fichero de registro de acceso (*access log*) del servidor en estudio. Se dispondrá de ficheros ASCII en los que se recogen datos de la actividad del usuario y de las peticiones realizadas. Cada registro del fichero contiene información de una solicitud de un documento. La información recogida se ajusta al formato *Commong Log Format - CLF* [2] a la que se ha incorporado el tiempo consumido en el servidor para procesar la petición. En este formato se recoge información básica de los accesos (peticiones) http al servidor y cada registro, línea del fichero, consta de los siguientes campos:

- cliente remoto: Dirección o nombre del cliente remoto que ha realizado la petición. Este dato ha sido anonimizado.
- Autentificación: Identificador de usuario remoto si se ha definido. Si no está definido se escribe un guión: **-**
- Autentificación: Nombre del usuario remoto si se ha definido. Si no está definido se escribe un guión: **-**
- Fecha y hora de la petición (Time Stamp). Formato: Fecha, hora y offset de la hora Greenwich (GMTx100)
- Petición (método y URL) enviada por el cliente:
  - GET – petición estándar de documento o programa
  - POST – Indica al servidor que a continuación vienen datos.
  - HEAD – Enlazar programas de comprobación y descargas
  - Otros: PUT, DELETE, TRACE y CONNECT
- status: Código numérico del resultado.Estado de la respuesta que proporciona el servidor
  - Éxito – series 200
  - Redirección – series 300
  - Fallo – series 400
  - Error servidor – series 500
- Volumen transferido en bytes. Tamaño en bytes del resultado (0 si no procede)
- *Elapsed time*. El tiempo consumido en servir la petición por el servidor (microsegundos)

Puede encontrar más información en la documentación de Apache [3].

En el procesamiento y valoración de la información recogida en el fichero original tenga en cuenta las siguientes consideraciones:

- No siempre los bytes transferidos coinciden con los bytes del tamaño del fichero. Esto puede ser debido, por ejemplo, a que el usuario aborta la conexión. En el log no es extraño que para el mismo recurso el nº de bytes transmitidos exhiba una cierta variabilidad. Por este motivo, y no teniendo información disponible sobre el tamaño real del recurso accedido, se tomará el valor medio de los volúmenes transferidos en bytes para ese recurso.
- En los ficheros de registro de la actividad hay datos de la actividad de clientes y usuarios del sitio web y de la actividad de los rastreadores, indexadores, etc (*bots, web crawlers, ...*). Este tipo de actividad no debería considerarse en el análisis y caracterización de la carga. Se puede diferenciar un acceso de usuario de uno de automático si en el *log* se recoge el agente, el identificador del browser utilizado por el usuario. En el *log* proporcionado para el desarrollo de esta práctica de laboratorio no se recoge esta información. En estos casos la diferenciación se puede hacer analizando la frecuencia y el tiempo transcurrido entre accesos, que en el caso de los *web crawler* sigue un patrón de alta frecuencia sin apenas diferencia de tiempo entre un acceso y el siguiente.
- Cuando se solicita una página web se solicitan todos los objetos que forman parte de esa página, fundamentalmente los elementos gráficos. En el registro de la actividad, el *log file*, se añadirá una línea por cada uno de los elementos solicitados por el brower y que juntos constituyen la página web requerida. Cada una de estas líneas se denomina *hit*. A la hora de realizar el análisis de la actividad del servidor web, todos estos *hits* que se derivan de la solicitud a una única página web han de agruparse dando lugar a lo que se denomina *page view*. Una serie de visitas a páginas de un sitio web por parte de un usuario pueden agruparse a su vez en una *sesión*. Esto permite una caracterización de la carga en base al análisis de las sesiones de usuario [4, cap. 6.2.4] y a un análisis de la actividad del usuario de gran interés para los departamentos de comercialización, publicidad y marketing de las organizaciones. Con los datos que proporciona un *access log* en el formato CLF, y no teniendo información del diseño del sitio web, la determinación de qué elementos componen una página puede ser establecida en base a la información de la ip del cliente remoto y al conjunto de elementos transferidos con la misma marca de tiempo. Para determinar las sesiones de usuario se suele considerar que un intervalo superior a los 30 minutos sin interacción con el sitio supone una nueva sesión de usuario.

Análisis de la utilización

El análisis de la utilización tiene implicaciones en la detección de problemas y ajuste del rendimiento del servidor. En este sentido el análisis de los datos de registro de accesos puede servir para abordar estudios relacionados con la decisión de qué documentos poner en caché, ajuste de sistemas relacionados con la gestión de la caché o el estudio del equilibrio entre peticiones y bytes.

Del análisis de los datos disponibles se puede conocer cuáles son los patrones temporales de accesos, qué recursos son los más populares y cuál es el origen y la frecuencia de las visitas. A partir de estas informaciones se obtiene una aproximación bastante precisa del tipo e intensidad de la actividad soportada por el servidor. En esta fase, por lo tanto, se elaborarán resúmenes y descripciones de:

- Los tipos y distribución de las respuestas del servidor: éxito, re-dirección o fallo.
- Las estadísticas globales de uso: Periodo de observación, hits, visitantes, ficheros descargados, etc.
- Tráfico soportado, bytes transferidos, patrones temporales por día de la semana y hora del día, evolución a lo largo del periodo de observación, etc.
- La popularidad de páginas y ficheros, descargas más frecuentes, accesos a directorios
- Errores en el servidor; Frecuencia y duración de las visitas

Respuestas del servidor

El análisis de los tipos de respuesta del servidor a una solicitud de servicio permite, por una parte, la identificación de las métricas a utilizar en el estudio de evaluación del rendimiento [5], y por otra,

poder evaluar la calidad del servicio (QoS) que proporciona dicho sistema. Las posibles respuestas que se pueden obtener de una petición de tipo *http* a un servidor web son las siguientes:

1. Éxito (successful)
    1. El usuario obtiene el documento solicitado. En este caso se consideran métricas de velocidad: tiempo de respuesta, productividad y utilización.
    2. No modificado. Ya tiene una copia del documento en caché. No se transfieren bytes.
2. Error - Documento encontrado (found) pero está en otro sitio
3. Error - No está, no tiene permiso o ha ocurrido algún error (en el servidor) que no permite acceder al documento.

Conocida la frecuencia y distribución de las diferentes respuestas del servidor el resto del estudio se centra en los casos de éxito. Estos serán, por regla general, los más frecuentes y los que provocan mayor carga en el servidor.

Análisis sobre los tipos y tamaños de los documentos

Desde el punto de vista de la demanda de recursos parece evidente que la solicitud de una página html estática que únicamente contenga texto utilizará muchos menos recursos que la solicitud de un video. En el análisis de la actividad del servidor web, y como una primera aproximación a la caracterización de la carga, se puede utilizar el atributo «*tipo de documento*» para la partición de la carga en clases homogéneas [4, cap. 6.3.5]

Para llevar a cabo este análisis deberá, en primer lugar, establecer las categorías de documentos (por ejemplo: Html, imagen, sonido, video, formateados, dinámicos,etc), a continuación deberá asociar cada documento a una de las categorías establecidas. Como no se tiene acceso a la información de diseño del sitio web, el tipo de documento se establecerá en base a la extensión del nombre de fichero. Puede que en el *log file* se encuentren extensiones desconocidas, en este caso se puede establecer una categoría *otros*.

Para cada clase de la partición definida por el tipo de documento se calcularán los valores que la caracterizan. Para cada tipo de documento se calcularán, al menos, los siguientes índices:

- Nº total de peticiones y porcentaje de accesos
- Media peticiones/día
- Peticiones distintas - Peticiones distintas/día
- Bytes totales transferidos (MB) - Media de MB transferidos / día
- Total bytes distintos transferidos (MB) - Total bytes distintos /día
- Media bytes transferidos (bytes)
- Mediana bytes transferidos
- Media de tamaño de fichero
- Mediana de tamaño de fichero.

Análisis sobre la popularidad de los documentos y transferencia.

Para ciertos tipos de estudios relacionados con la evaluación de rendimiento en sistemas web es conveniente realizar el análisis y caracterización de la demanda de los documentos individuales del sitio web.

Se estudia la popularidad de los documentos y el volumen de transferencia de bytes que esto implica. Este tipo de estudio permitirá determinar, por ejemplo, qué documentos concentran el mayor porcentaje de peticiones y bytes. Información que puede ser utilizada por el administrador del sistema para determinar qué documentos se pueden poner en caché de servidor, de red o de cliente para mejorar el rendimiento del servicio.

Para realizar el análisis cada documento se identificará por su *uri* completo. Esto implica que si un documento se ha movido a otro directorio se considerará un documento diferente. Se estudian las

relaciones entre los accesos al sitio web, el tamaño total (en bytes) transferido y el nº de documentos diferentes accedidos. Se elaborarán, al menos, los siguientes índices:

1. Peticiones documentos diferentes / Peticiones totales
2. Bytes diferentes / bytes totales
3. Ficheros diferentes accedidos solamente una vez
4. Bytes diferentes accedidos solamente una vez.
5. Distribución de los tamaños de los ficheros (frecuencia acumulada).
6. Comportamiento de la referenciación de ficheros.
    1. Concentración de referencias - popularidad : acumulación de referencias independientemente del tiempo.
        1. Clasificación descendente de los ficheros siguiendo el criterio del nº de veces que han sido solicitados.
        2. Distribución por accesos, frecuencia acumulada.

Construcción del modelo de carga

El proceso de construcción de la carga comprende una serie de fases que van desde el filtrado y depuración de los datos de actividad hasta la asignación de valores a los parámetros del modelo [1]. Sin embargo, en esta práctica no se abordará la asignación de valores a los parámetros del modelo.

Filtrado y depuración de datos.

Para la construcción del modelo de carga se van a considerar únicamente las peticiones que tienen una cierta contribución a la carga del sistema. En el caso de los servidores estáticos con información de recuperación y consulta, las peticiones a considerar han se ser las peticiones de éxito de tipo *GET*.

Si se opta por el estudio del fichero completo, es posible que se tenga que partir el fichero para el análisis independiente de cada parte. En estos casos lo habitual es efectuar una partición siguiendo un criterio temporal (semana, día, ...).

En casos de conjuntos voluminosos de datos, se puede optar por realizar la caracterización de la carga sobre una muestra de datos. En este caso se aplicarán las técnicas de muestreo que garanticen un muestra representativa del conjunto completo de datos.

En resumen, para la caracterización de la carga se tendrán en cuenta las siguientes consideraciones y restricciones:

- Población: Documentos solicitados al servidor web.
- Componente carga: Transacción HTTP GET
- Criterio de selección de componentes: Código de respuesta 2xx o 3xx
- Parámetros: Tamaño (medio) del fichero, nº de accesos al fichero y tiempo de ejecución.

Partición de la carga

Como se ha visto en el tema de carga de trabajo, es necesario dividir el conjunto de componentes de la carga en subconjuntos homogéneos y compactos para mejorar el grado de similitud entre el modelo y la carga real. Establecida la partición, se elige un representante de cada subconjunto y se construye el modelo a partir del representante o representantes de cada partición. A cada representante o representantes de la partición se les asignará un peso en función de la importancia relativa de cada clase, partición, en la población total.

Para el tipo y objetivos del estudio propuesto, se puede comenzar con una partición por atributos. Esta aproximación, al no tener en cuenta otros factores que intervienen en el rendimiento, se reforzará o sustituirá por una clasificación en base a la aplicación de las técnicas de *clustering*.

Informe

Una vez finalizado el análisis y la caracterización de la carga se elaborará un informe por grupo. Incluirá, al menos, los contenidos que se indican a continuación. No obstante, si el grupo lo cree conveniente, se pueden añadir secciones y/o anexos.

1. Presentación del documento y del estudio. Objetivos y alcance del estudio.
2. Descripción del sistema en estudio (SUT), del componente de carga y de los métricas.
3. Información disponible. Descripción de las características de los datos y de las informaciones recogidas en el fichero *log*.
4. Análisis de los datos. Se espera un análisis y discusión acerca de, al menos, los siguientes aspectos del servidor web:
   - Estadísticas globales de uso: Periodo de observación, hits, visitantes, ficheros descargados, ...
   - Popularidad de páginas y ficheros, descargas más frecuentes.
   - Visitantes: Frecuencia y duración de las visitas
   - Patrones temporales del tráfico soportado por el servidor en términos de accesos y volumen (bytes) transferidos.
   - Caracterización de las respuestas del servidor, de los tamaños (en Kb) de la respuesta y del throughput del servidor.
5. Caracterización de la carga. Partición y modelo.
6. Conclusiones
7. Bibliografía
8. Anexos con los procedimientos o scripts para el tratamiento y análisis de los datos.
9. Anexos con las tablas y contenidos que dan soporte a los contenidos del informe.

El documento final ha de ajustarse a las normas de estilo elaboradas para las prácticas de laboratorio de esta asignatura.

Bibliografía

[1] X. Molero, C. Juiz, and M. Rodeño, *Evaluación y Modelado del Rendimiento de los Sistemas Informáticos.* Pearson Educación, S.A., 2004. [Online]. Available: http://www.pearsoneducacion.com/molero/

[2] Logging in w3c httpd. [Online]. Available: https://www.w3.org/Daemon/User/Config/Logging.html#common-logfile-format

[3] Log files - apache HTTP server version 2.5. [Online]. Available: https://httpd.apache.org/docs/trunk/logs.html#common

[4] D. A. Menasce and V. A. F. Almeida, *Capacity Planning for Web Services: Metrics, Models, and Methods.* Prentice Hall, 2001, vol. 1 edition.

[5] R. K. Jain, *The Art of Computer Systems Performance Analysis: Techniques for Experimental Design, Measurement, Simulation, and Modeling.* Wiley, 1991, vol. 1 edition.