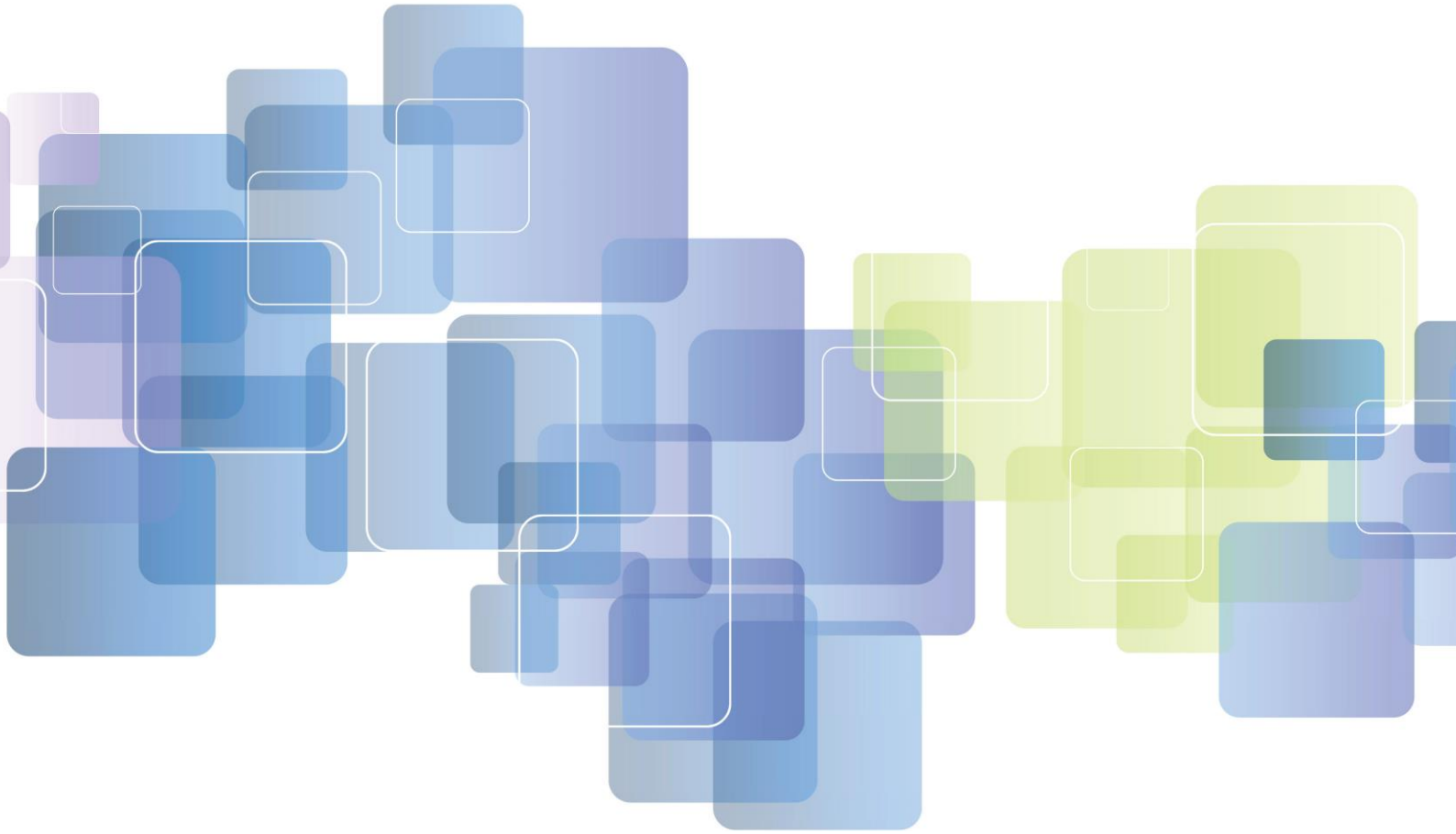
TecnetDati

Monitoring end-to-end delle applicazioni Web

Giorgio Marras e Paolo Cremonesi



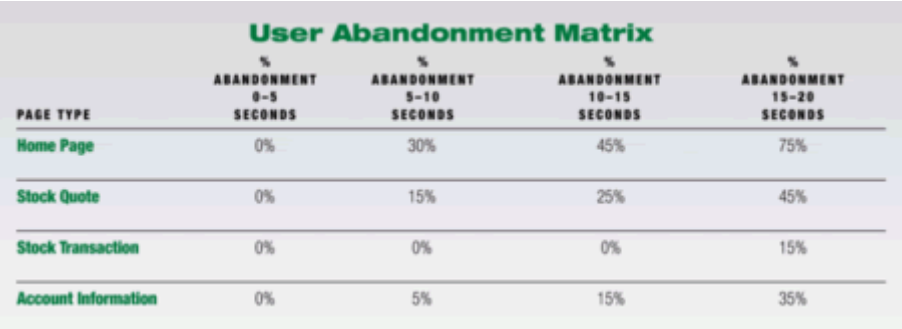
Misurare i tempi di navigazione di un'applicazione Web come li percepisce l'utenti finale

Lo scopo della rilevazione delle prestazioni end-to-end delle applicazioni Web è quello di misurare i tempi che riflettono il comportamento di un utente Web durante la navigazione.

I campi di applicazione di monitoring end-to-end sono molteplici, per esempio:

- misurare la qualità del servizio erogato agli utenti
- individuare quali sono i componenti dell'applicazione Web responsabili di un eventuale malfunzionamento o comunque suscettibili di miglioramento.

Le statistiche mostrano che vi è una correlazione stretta tra prestazioni delle applicazioni Web e tasso di abbandono degli utenti (vedi figura 1).



User Abandonment Matrix

PAGE TYPE	% ABANDONMENT 0–5 SECONDS	% ABANDONMENT 5–10 SECONDS	% ABANDONMENT 10–15 SECONDS	% ABANDONMENT 15–20 SECONDS
Home Page	0%	30%	45%	75%
Stock Quote	0%	15%	25%	45%
Stock Transaction	0%	0%	0%	15%
Account Information	0%	5%	15%	35%

figura 1 - tempo di risposta vs. tasso di abbandono

Perché il monitoring sia efficacie, è importante misurare grandezze di dettaglio e grandezze di alto livello. Il tempo di risposta di una pagina non è di per sè interessante se non correlato ad una operazione di business. Per un'azienda di credito, per esempio, non è interessante sapere che la pagina "xxx" risponde mediamente in n secondi, ma piuttosto che l'operazione di bonifico richiede mediamente n secondi per essere portata a termine.
Le considerazioni che seguono vogliono dare una rassegna breve ma esaustiva sull'argomento.

La pagina Web

Una pagina Web è un documento scritto in linguaggio HTML ed è composto da un insieme di oggetti che possono essere immagini, dati multimediali, altri documenti HTML, ecc... La pagina stessa è un oggetto (l'oggetto contenitore). Per ciascuno degli oggetti di una

pagina (compreso l'oggetto contenitore) il browser effettua una serie di operazioni caratteristiche del protocollo http (Hyper Text Transfer Protocol, si veda la figura 2):

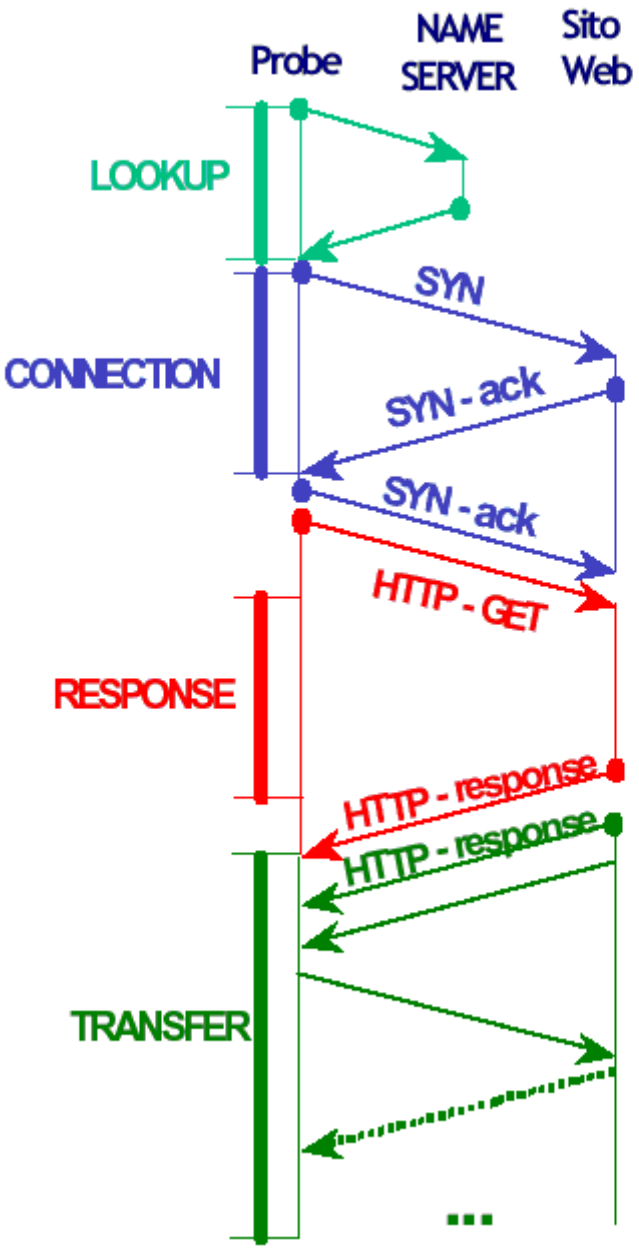


figura 2 - successione delle operazioni http per la richiesta di un oggetto

- a partire dall'URL dell'oggetto occorre effettuare una prima fase di lookup che consente di trasformare l'indirizzo simbolico del server a cui richiedere l'oggetto (ad esempio, http://www.neptuny.it/) e di ottenere l'indirizzo IP (l'indirizzo numerico) del server stesso. L'operazione di lookup consiste nel contattare un server DNS ed effettuare una richiesta di risoluzione di indirizzo. I browser utilizzano una cache di indirizzi IP per cui la risoluzione vera e propria è fatta una

sola volta all'interno di una transazione. Il tempo necessario per eseguire il look up è chiamato lookup time
- effettuata la fase di lookup, per ottenere l'oggetto è necessario aprire una connessione TCP con il server (il tempo necessario per aprirla è detto connection time). I browser possono aprire più connessioni contemporaneamente per scaricare gli oggetti di una pagina. Internet Explorer, per esempio, utilizza al più 4 connessioni contemporanee al server
- ottenuta la connessione, il browser richiede l'oggetto al server con l'invio di una richiesta http, alla quale il server risponde con una risposta http. Il tempo tra l'invio della richiesta e l'arrivo del primo byte della risposta del server è definito response time e rappresenta un tempo proporzionale alla velocità del server nel soddisfare le richieste
- una volta ricevuto il primo byte, l'oggetto è inviato al client con un numero di pacchetti dipendente dalla sua dimensione. Il tempo necessario per questa fase è detto download time.

Il tempo totale di oggetto è la somma delle componenti indicate.
Il tempo totale di pagina è il tempo tra l'inizio dello scaricamento del primo oggetto (il documento HTML della pagina) e la fine dello scaricamento dell'ultimo oggetto contenuto nella pagina.

Per poter determinare il peso relativo delle singole componenti (lookup,connessione, risposta, download) di una pagina, è necessario pesare in modo opportuno le componenti di tutti gli oggetti che compongono la pagina.
Si definisce, ad esempio, tempo di risposta della pagina il valore:

$$T_R = \frac{\sum_{\forall i} T_R^i}{\sum_{\forall i} T^i} \cdot T$$

ove sia:

T tempo totale di pagina

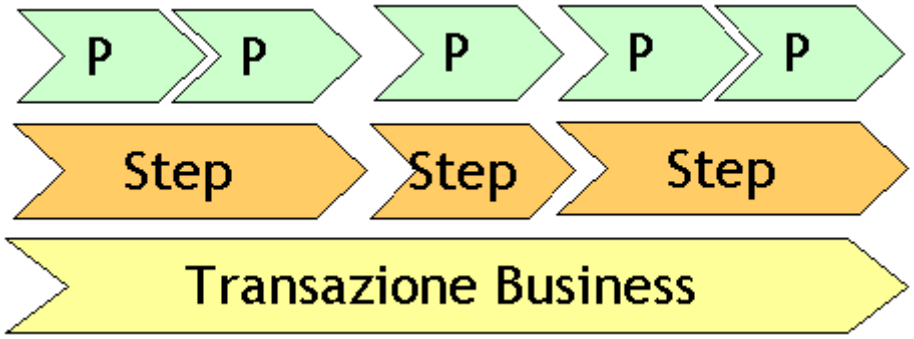
T_R^i tempo di risposta dell'oggetto i-esimo

T^i tempo totale dell'oggetto i-esimo della pagina

In pratica ciò corrisponde ad assegnare alla pagina un tempo di risposta proporzionale al "peso" che la fase di risposta ha avuto nello scaricamento di ciascun oggetto. Analogamente si procede per i tempi di lookup, di connessione, di download.

Transazione Web e operazioni di business

Possiamo definire una transazione Web come una successione di azioni (le azioni possono essere, ad esempio, una sequenza di click sui link di un sito Web), legate tra loro da caratteristiche omogenee rispetto alle operazioni da eseguire. Tali operazioni devono produrre un effetto legato al business dell'azienda e sono comunemente definite *operazioni di business*. Nel caso specifico di un azienda di credito saranno operazioni di business, ad esempio: Movimenti c/c, Bonifici, Pagamento tasse F24, Operazioni su Fondi.
Passando da un livello all'altro (dall'azione all'operazione), il *tempo di risposta* si ottiene come somma dei tempi rilevati. Ad esempio, il tempo di risposta di una pagina può essere misurato mediante strumenti specifici, mentre il tempo totale di una transazione Web è la somma dei tempi di ciascuna pagina che compone la transazione. Misurare il tempo di risposta di una transazione web richiede pertanto la definizione di regole di mapping, cioè regole che permettono di associare tra loro le pagine e le transazioni. Nella pratica, per descrivere in modo compiuto un'operazione da misurare, può essere utile introdurre un ulteriore livello intermedio, tra la transazione e la pagina, costituito dallo step (o anche passo di navigazione). Uno step è una sequenza ordinata di azioni che un utente deve seguire per poter compiere una transizione. Spesso ad ogni step corrisponde una ben precisa pagina Web.
Ma possono esserci casi in cui ad uno step corrispondono due o più pagine (per esempio, nel caso in cui ad un click su un link segue l'apertura di due finestre).



Esempio: Transazione "Bonifico"

üStep 1: Homepage

üStep 2: Login

üPagina 2.1: Click su Login

üPagina 2.2: Submit form autenticazione

üStep 3: Click su Pagamenti

üStep 4: Submit modulo

üStep 5: Conferma bonifico

Per le operazioni di business, inoltre, vi è un'altra grandezza importante da rilevare per avere una prima rappresentazione del livello di servizio erogato: la *disponibilità*, cioè la percentuale di operazioni eseguite con successo. La disponibilità misura la percentuale di transazioni che sono andate a buon fine, ossia in cui è stato possibile eseguire tutti gli step senza errori. Una transazione può fallire quando uno dei suoi step genera un errore, ad esempio quando il sistema è troppo carico, o quando si ha il down di un server, oppure quando vi sono errori nell'applicazione Web.

Misure attive e passive

Il tempo di risposta di una transazione web può essere misurato in due modalità, tra loro complementari:

- rilevazione attiva, effettuata mediante sonde "Agents" costituite da "robot" Web che riproducono (simulano) le transazioni da misurare ad intervalli regolari (campionamento temporale) e ne rilevano le prestazioni
- rilevazione passiva che misura il tempo effettivo delle transazioni percepite dagli utenti reali (traffico reale). Il traffico reale del sito viene monitorato mediante sniffing del traffico di rete (sonde "Probes")

Le sonde "Agent" permettono di definire a priori la suddivisione in step/pagine delle transazioni da misurare. La rilevazione attiva prevede un campionamento anche nello spazio: occorre infatti scegliere alcuni *punti di vista* da cui effettuare la rilevazione. In funzione degli obiettivi si possono scegliere come punti di vista la Rete interna dell'azienda (dove sono collocati gli Application Server) o la rete pubblica (attivando rilevazioni mediante Internet Provider).
Collocando opportunamente le sonde e attivando un campionamento regolare si può pertanto:

- rilevare il livello di servizio del sistema, cioè misurare l'efficienza dell'applicazione e dell'infrastruttura tecnologica dell'azienda
- confrontare il comportamento di differenti provider, tecnologie, eccetera.

La rilevazione passiva, mediante sonde "Probes", prevede che le grandezze siano campionate sulla base delle richieste reali al sito (ovvero nei tempi e dalle posizioni da cui provengono le richieste reali). La provenienza della richiesta definisce il concetto di *dominio chiamante*. Classificando gli indirizzi IP delle richieste ai Web Server, è possibile classificare il traffico in base alla provenienza, per esempio, identificando in modo distinto le transazioni provenienti dalle diverse filiali di un'azienda di credito. Le sonde "Probes" rilevano il tempo di risposta delle pagine configurate e calcolano il tempo complessivo della transazione sulla base delle regole di mapping di cui abbiamo parlato in precedenza.

Occorre osservare che il tempo di risposta di una pagina Web è una grandezza con distribuzione statistica a coda lunga (heavy-tailed) a causa dei fenomeni legati al comportamento della rete. Nel caso di questo tipo di grandezze, il descrittore statistico "media" non è sufficiente a caratterizzare la grandezza in quanto la varianza è troppo elevata; è necessario pertanto utilizzare descrittori più significativi come i percentili

(per esempio 90 o 95-percentile) e le mediane.

A proposito di mediane e di tempi di oggetto, si noti che rispetto alle mediane può accadere che la mediana del tempo totale di un oggetto non coincide con la somma delle mediane delle singole componenti.

Aggregazioni e criteri di campionamento

In un contesto complesso come quello di un'azienda di credito è fondamentale gestire la complessità e i volumi mediante tecniche statistiche, per esempio *campionamento e aggregazione*, al fine di rendere gestibile il processo di elaborazione dei dati misurati dagli strumenti di monitoring. In termini più generali, l'obiettivo è di raccogliere tali dati, elaborarli opportunamente e archiviarli in un data warehouse tematico che chiameremo PERFORMANCE WAREHOUSE (PWH).
Il processo di trattamento dei dati contenuti nel PWH prevede fasi distinte: *Misura e Aggregazione*.
Nella prima fase (Misura) vengono attuate tecniche di campionamento, cioè tecniche statistiche mediante le quali viene rilevata solo una parte del traffico, scelta in modo da fornire una rappresentazione probante dell'intero campione che si sta esaminando. In altri termini è necessario stabilire quali sono le transazioni da esaminare e qual è la frequenza con cui le grandezze vengono misurate.

Nella seconda fase (Aggregazione) le grandezze rilevate sono aggregate sulla base di diversi concetti, per esempio la dimensione temporale (aggregazione temporale). In tal caso i dati rilevati secondo la risoluzione di campionamento sono aggregati ad un orizzonte temporale meno granulare, per esempio ora, giorno, mese.
Ai diversi livelli di aggregazione vengono calcolate statistiche sui dati rappresentate mediante i descrittori statistici media, percentile, eccetera.

Oltre ad utilizzare le aggregazioni temporali, può essere opportuno distinguere le differenti fasce orarie di funzionamento dei sistemi. Riportiamo di seguito le tre fasce orarie identificate in un recente caso utente e precisamente:

- Ore lavorative (L)
- Ore non lavorative (NL)
- Ora (Ore) di punta (P), definisce (definiscono) un intervallo temporale critico per il carico dei sistemi.

In tal caso l'indicazione della fascia oraria affiancherà quelle dell'aggregazione temporale. In pratica se si scrive Giorno-L si intende ad esempio la media giornaliera sulle ore giorni lavorative, e con Giorno-P la media giornaliera nelle ore di punta, mentre Mese-L indicherà la media mensile nelle ore lavorative, e così via.

Casi di successo

Tecnet Dati e Neptuny collaborano con alcune delle principali aziende italiane dei settori Credito e Telecomunicazioni alla realizzazione di progetti per la rilevazione dei tempi di risposta end-to-end delle applicazioni Web (Monitoring end-to-end) e la pianificazione del carico dei sistemi (Capacity Planning).
Per saperne di più potete inviare una mail a paolo.cremonesi@polimi.it o giorgio.marras@tecnetdati.it.

Tecnet Dati s.r.l.
C.so Svizzera 185 –
10149 - Torino (TO), Italia
Tel.: +39 011 7718090 Fax.: +39 011 7718092
P.I. 05793500017 C.F. 09205650154
www.tecnetdati.com