

TECNOLOGIE INFORMATICHE PER IL WEB

Federico Mainetti Gambera

28 marzo 2020

Indice

1	Calendario delle lezioni	2
2	Introduzione al corso	2
3	Architetture	2
4	HTTP	3
5	CGI, Common Gateway Interface	3
6	Servlet - base	4
7	Servlet - esempi	6
7.1	Esempio: contatore condiviso fra i client	6
7.2	Esempio: contatore per ogni client	6
7.3	Forme di identificazione	6
7.4	Tracciamento della sessione	7
8	JDBC	9

1 Calendario delle lezioni

<https://docs.google.com/spreadsheets/d/14ShTdkFCZ63MlyKP0LCSLPsPAeiu8D2sYVXSrkw9B6k/edit#gid=0>

Prerequisiti

Introduzione al corso - Evoluzione architetture
Protocollo HTTP
HTML 3 e 4
CGI

2 Introduzione al corso

Non ci saranno prove in itinere per questo corso.

E' opportuno guardare i video relativi agli argomenti indicati prima di partecipare alla lezione.

Queste lezioni hanno lo scopo di aggiungere informazioni, chiarire dubbi e aiutarci a capire quali siano gli argomenti più importanti e i punti chiave.

I materiali video coprono interamente gli argomenti del corso.

3 Architetture

Il web è una piattaforma per sviluppo di applicazioni con un architettura molto particolare. Per architettura si intende l'insieme delle risorse (hardware, connettività, software di base, software applicativo). Le applicazioni web hanno una particolare conformazione della loro architettura che prevede tre livelli: Client, Middle-tier, Data-tier.

Le architetture sono cambiate molto negli anni. Se ne individuano tre grandi famiglie: Mainframe, Client-Server, Multi-tier (per esempio le applicazioni web).

Il compito di questo corso è quello di riuscire a programmare nel Middle-tier con qualche accenno al Client. In un'applicazione web il Client interagisce con il Middle-tier con un preciso protocollo, che non è il classico tcp-ip, ma il protocollo HTTP.

[Una lettura molto importante per il corso è il documento RequestforComment 1945, Tim Berners Lee (<https://www.w3.org/Protocols/rfc1945/rfc1945>), che rappresenta l'atto di fondazione del web.]

Il Client manda delle request al Server che invia delle responses. Le request del Client sono gestite tramite un'applicazione detta User agent (browser).

Http è rivoluzionario per la sua semplicità: le richieste del Client e le risposte del server sono delle semplici stringhe.

Architettura delle applicazioni web: c'è un client (pc) con un user agent (browser) che emette richieste che vengono servite con delle risposte da un web server, detto anche HTTP server. L'architettura di cui però ci occupiamo è più complicata di così, perchè vede alle spalle del web server un application server (TomCat), che ha lo scopo di calcolare una risposta "personalizzata" secondo parametri e criteri precisi e di produrre una pagina web appropriata.

[Installare un'architettura completa sul pc che prevede tutti i livelli appena visti (guida nella cartella strumenti) (si può usare anche IntelliJ invece di Eclipse)]

Altri elementi dell'architettura sono il proxy e il gateway. Il proxy è un intermediario fra un client e un origin server, per definizione può comportarsi sia come server sia come client. Fa copie di risorse e le inoltra ai client e può essere anche utile per limitare gli accessi e gestire autorizzazioni. Il Gateway serve per far interagire un'applicazione web con applicazioni che non usano il protocollo HTTP, quindi il gateway traduce richieste HTTP per applicazioni che non lo comprendono, tipicamente usato per SQL.

4 HTTP

Come si identifica una risorsa? con l'URL (Uniform Resource Locator) che è una stringa formattata in maniera molto semplice: prefisso (protocollo), indicazione della macchina fisica in cui è presente la risorsa, una eventuale porta in cui la macchina ascolta le richieste e un Path. Gli URL sono estremamente semplici, e utilizzano concetti già noti in precedenza.

HTTP request: è una banale stringa che contiene una request-line, degli header e un allegato (per gli upload) facoltativi. la request-line contiene tre informazioni: il metodo (funzione) della richiesta, l'URL, la versione del protocollo usata. I metodi di richiesta sono principalmente due, GET e POST; questi metodi possono essere visti come chiamate di funzione.

HTTP response: anche questa è una banale stringa, che, ancora più semplicemente contiene un codice di stato (es. 404 file not found), degli header facoltativi, e un allegato.

La conseguenza di un protocollo così semplice è che con un solo client si può interagire con tutti i back end. La complessità si sposta però nell'application server.

Gli header sono informazioni opzionali a cura del browser, trasmesse come fossero parametri che aggiungono informazioni alla request o alla response.

5 CGI, Common Gateway Interface

Tecnologia ormai morta e superata.

CGI è una convenzione che standardizza un certo numero di variabili d'ambiente (condivisibili fra più processi) grazie alle quali il web server può smontare una richiesta HTTP e salvarla in una zona di memoria alla quale un processo di un'applicazione esterna può accedere.

Per invocare un'applicazione tramite CGI, l'URL della richiesta HTTP deve presentare un path che porta a un'applicazione eseguibile. Una volta ricevuta una richiesta HTTP che rimanda a un eseguibile, il web server smonta la richiesta, ne preleva le informazioni (parametri) e le salva nelle variabili d'ambiente, successivamente fa eseguire all'application server il programma indicato nel path dell'URL. L'application server preleva le informazioni dalle variabili d'ambiente e costruisce un file HTML che poi verrà mandato come risposta al client.

CGI ha due grandi difetti, il primo riguarda la sicurezza, in quanto i file eseguibili erano direttamente accessibili, il secondo è che siccome i processi dell'application server muoiono dopo ogni esecuzione, non c'è modo di creare un sistema che mantiene una sessione attiva. Inoltre le prestazioni di CGI sono molto basse, per esempio, siccome i processi muoiono continuamente, c'è un continuo bisogno di stabilire connessioni coi database, che è un'operazione onerosa.

Prerequisiti

Fondamenti Servlet - base

6 Servlet - base

Nasce l'esigenza di generare contenuti dinamici per le applicazioni web, inizialmente, infatti, HTTP era concepito come protocollo per scambio di documenti.

Il gateway è un elemento imprescindibile che aumenta le capacità di un'applicazione web. Abbiamo visto che la versione arcaica di gateway era il CGI, che rappresenta il modo più semplice e che usa le variabili d'ambiente per adempiere al suo compito. Abbiamo però visto che CGI ha diversi problemi, perciò è necessario trovare una nuova soluzione.

Vediamo ora il meccanismo più popolare (almeno fino a qualche anno fa) per realizzare i requisiti di produzione dinamica di contenuti. HTTP è nato per essere semplice, ha solo due metodi, GET e POST, non prevede l'identificazione di un client, tutte le richieste sono uguali. Per un Web Server la nozione di sessione non esiste. I metodi GET e POST sono richieste parametriche, cioè alle quali si possono aggiungere informazioni sotto forma di parametri stringa. Il metodo GET i parametri li mette nella query-string, il metodo POST li mette nel body. HTML ha un costrutto "form" che serve al browser per costruire un richiesta dove i parametri sono complicati, per esempio richieste dove un parametro è un file. Il metodo POST è quindi spesso usato con il costrutto HTML form.

Java Servlet è un modo nuovo, rispetto a CGI, di strutturare l'applicazione che riceve la request e formula la response.

Al programmatore di un Servlet si chiede di programmare una classe, egli dovrà lavorare all'interno di un Framework, cioè una soluzione parziale a una serie di problemi con caratteristiche comuni. Il framework è uno schema, una soluzione parziale che omette la parte variabile di una serie di applicazioni con qualcosa in comune. Java Servlet è appunto un framework, tutte le applicazioni web hanno in comune tutto il protocollo HTTP. Dal framework ci aspettiamo quindi di non dover programmare la gestione delle request e delle response.

Il Servlet container è un ambiente che esegue il tuo programma, nel nostro caso è Tomcat. Il container materializza l'API del framework che stiamo usando, per esempio il metodo doGet() di cui noi facciamo l'override viene chiamato dal Servlet container.

Alle spalle del web server dunque sta la JVM, all'interno del quale risiede il servlet container che gestisce i Servlet che il programmatore scrive.

Un primo beneficio fra che si nota nel programmare in un ambiente così fortemente strutturato è che, diversamente da CGI, i Servlet vengono eseguiti all'interno di un ambiente persistente e quindi può eseguire più di una richiesta senza essere terminato. Non c'è bisogno di preoccuparsi della concorrenza, semplicemente si programma un Servlet pensando a come deve rispondere a una certa request. Siamo quindi in presenza di un ambiente stateful. C'è un prezzo da pagare per questo beneficio, ovvero che non siamo noi a gestire la concorrenza e quindi i thread, il modello di concorrenza lo ereditiamo dal contenitore. Il modello da seguire è chiamato "one thread per request" e significa che una volta sviluppato un servlet non siamo noi a invocare la "new" su quell'oggetto, perchè è il contenitore a gestire questo aspetto e lui ne avrà sempre e soltanto uno istanziato. Tutte le request interagiranno sempre con lo stesso oggetto istanza della Servlet programmata da noi. Quindi, si programma una servlet, ce ne sarà una sola istanza, un solo oggetto, tutte le request che riceviamo (tante nello stesso istante) avrà un thread allcoato, non da noi, ma dal container e questi thread agiranno tutti sulla stessa istanza del Servlet. Questo processo ha ovviamente un grande problema: le variabili della classe Servlet sono condivise per tutte le request.

Un framework ti dà servizi, ma ti impone dei vincoli.

Dal punto di vista materiale un framework è una libreria e noi in particolare useremo javax.servlet e javax.servlet.http, che sono interfacce implementate dal contenitore.

Il contenitore mappa le request sui metodi doGet() e doPost().

Il metodo destroy() viene chiamato quando il processo dell'applicazione viene stoppato, fatto che è deciso dall'amministratore del server.

Il file web.xml serve per mappare le richieste http al corretto servlet.

La programmazione per Servlet è una programmazione per componenti, nel senso che il programmatore scrive solo i componenti variabili per l'applicazione, tutto il resto è gestito dal framework.

Il servlet Context è una scatola che contiene diversi componenti che presi nell'insieme rappresentano un'applicazione. L'oggetto java che rappresenta un'applicazione è proprio il servlet context, infatti quando devo fare la deploy di una applicazione si distribuisce il servlet context. Il servlet context è un insieme di risorse che contiene i sorgenti delle applicazioni java scritte dal programmatore, i file di configurazione e le risorse stesse (per esempio le immagini).

Il file web.xml contiene la configurazione dell'applicazione, fra cui la mappatura delle varie servlet. Per maggiori informazioni guardare i video. In breve: c'è Tomcat che è rappresentato dal server stesso, il contenitore, nei nostri esempi sarà localhost:8080, a Tomcat arriva una richiesta. Una richiesta appende all'indicazione del server (localhost:8080) un url, che è quello che consente di scatenare la servlet corretta. Tomcat viene istruito tramite un processo di Servlet mapping. Per maggiori informazioni guardare i video o i lucidi.

Prerequisiti

Fondamenti Servlet - esempi

Informazioni

Su Beep nella sezione Documents and media > Esercizi > Servlet - jsp - jdbc > esercizio server side messaggi, si trova un esempio (progetto eclipse completo .zip) commentato con video e power point. E' importante, dice il prof, perchè presenta tutto ciò che può esserci in esame.

Sempre su Beep, ma nella sezione Documents and media > Esercizi > CSS c'è un esercizio, che però è integralmente spiegato nella cartella delle video lezioni riguardanti Css.

Useremo i giorni di sospensione (8-9 aprile) per le prove in itinere per fare degli esercizi autonomi.

Verranno pubblicati anche altri progetti, man mano più complessi.

Sono molto importanti la documentazione sul package `javax.servlet.http` e il package `javax.servlet`.

7 Servlet - esempi

Una programmazione in servlet è una programmazione che ci chiede di rispettare alcune regole per trarne dei benefici. I benefici principali sono la scalabilità, la gestione automatica dei cicli di vita, la mascheratura degli aspetti tecnici di HTTP. Java servlet ci permette di interagire con HTTP attraverso degli oggetti. Uno dei primo esempio di oggetto che si incontra è l'oggetto request.

7.1 Esempio: contatore condiviso fra i client

Il file `web.xml` ci permette di controllare le impostazioni con cui vogliamo lavorare, inoltre ci permette di definire parametri a livello di applicazione (globali) o parametri a livello di servlet.

I parametri possono poi essere reperiti con chiamate di metodi di oggetti di sistema in maniera molto semplice.

Con questo esempio, oltre a mostrare l'utilizzo di questi parametri, viene mostrata l'utilità dei dati membro all'interno della servlet (gli attributi dell'oggetto servlet). Esistendo una sola istanza di servlet per tutte le request, ci aspettiamo che i dati membro vengano usati in maniera concorrente per tutti i client.

Nell'esempio particolare del contatore possiamo vedere che tutti i client possono incrementare il valore del contatore, perchè condividono il dato membro della servlet.

Quindi con questo esempio si imparano due cose importanti: come si configura una servlet per farla partire e il particolare modello di gestione della concorrenza e delle variabili.

7.2 Esempio: contatore per ogni client

Con questo esempio andiamo ad analizzare il meccanismo di una sessione. Per sessione si intende un gruppo di richieste che possono essere fatte risalire a un unico cliente. HTTP ha solo il metodo GET e POST, e quindi sembra che non sia lui ad occuparsi della gestione della sessione, infatti la gestione della sessione non era nativamente intesa all'interno del protocollo HTTP, ma si è sviluppata successivamente grazie all'utilizzo degli header, in particolare esistono due tecniche: o l'uso degli header cookie, oppure grazie all'url rewriting.

Da un punto di vista programmatico queste due tecniche sono trasparenti per noi programmatori, noi otterremo un oggetto che maschererà i meccanismi che stanno dietro alla gestione delle sessioni.

7.3 Forme di identificazione

Il protocollo HTTP nella sua versione più base (senza i metadati aggiuntivi) serve richieste anonime. [tema molto importante per il professore, spesso usato all'orale:] Analizziamo la terminologia e i concetti

base che stanno dietro alla sicurezza della gestione delle sessioni (con forme di identificazione del client via via più stringenti):

- Pseudo identificazione: l'atto con cui il server associa un'etichetta arbitraria alle richieste del client allo scopo di identificare quelle che provengono dallo stesso client. Come esempio si può vedere in atto la pseudo identificazione si può usare il "SessionCounter", aprire due browser (uno in incognito e uno no) e vedere come i due contatori non sono condivisi. La pseudo identificazione non ha bisogno che l'utente si logghi.
- Identificazione: è il processo con cui il cliente dichiara un'identità. Non ci stiamo più riferendo a un client, ma a un user, un account, una persona.
- Autenticazione: è il processo in cui il server verifica l'identificazione del client. Il client e il server condividono un "segreto" (password). La differenza fra pseudo identificazione e identificazione-autenticazione è che nel primo il server inventa un label da dare a un client, che comunque rimane anonimo in quanto l'identità è fornita dal server, e nel secondo il client si dichiara un user che poi viene verificato dal server, quindi in questo caso l'identità proviene dal client.
- Autorizzazione: è il processo in cui il server garantisce l'accesso a risorse e processi a un utente identificato e autenticato. L'autorizzazione è una proprietà dell'identità verificata. E' in poche parole il permesso di vedere o fare cose particolari a fronte di un'identità verificata.
- RBAC (Role based access control): è uno schema di autorizzazione che garantisce diritti non solo sulla base della tua identità, ma anche grazie al tuo ruolo.

7.4 Tracciamento della sessione

Si dice cookie un piccolo contenuto di informazioni che il server installa sul browser del client, allo scopo che venga restituito al server a ogni successiva richiesta per poter tracciare un client e la sua sessione.

Vediamo ora in maniera "tecnica" come funziona la pseudo-identificazione tramite gli header cookie. Il server, quando riceve la prima richiesta da parte di un client (quindi una richiesta anonima), vuole che la seconda richiesta sia pseudo-identificata (non più anonima). Se andiamo ad analizzare la seconda richiesta infatti ci accorgiamo che fra i request header è presente, nella voce cookie, il label che il server ha generato per me. Infatti se andiamo a vedere la response alla prima request, notiamo che il server propone al client un label tramite l'header set-cookie. Dunque alla seconda richiesta il client restituisce al server questo cookie e quindi riesce a pseudo-identificarsi.

Fra i vari esercizi caricati su beep ce n'è uno che si chiama cookie inspector, con cui possiamo andare a fondo con la questione dei cookie. Come esercizio provare anche a disabilitare i cookie e a vedere cosa succede.

I cookie non sono un canale sicuro, possono essere sniffati e si possono fare attacchi man in the middle.

L'identificazione autorizzata non sostituisce la pseudoidentificazione. Immaginiamo che avvenga sia una pseudoidentificazione sia una identificazione autenticazione con per esempio un login, quindi il mio client è stato autenticato e la sessione è tracciata dai cookie. Da ora in poi è ovvio che la sola pseudoidentificazione non implica l'autenticazione precedentemente fatta (altrimenti dei semplici attacchi informatici potrebbero causare grandi problemi di falsa appropriazione di identità), ovvero il tracciamento della sessione e l'identificazione autenticazione sono su due piani diversi.

Il session tracking implica la pseudoidentificazione del client, o con cookie, dove il server inietta una stringa e il browser la restituisce a ogni futura richiesta, o con url rewriting, ormai non più usato.

Invece allo scopo di associare allo pseudo-cliente una qualunque informazione durevole (per esempio l'autenticazione di un client, oppure l'inserimento di un prodotto in un carrello), il server abbina a ogni pseudoidentità un oggetto Java chiamato Session che è controllabile dal programmatore della servlet. In questo oggetto può essere presente un'informazione che conferma che questa pseudoidentità è autenticata e corrisponde a un determinato user. In questo oggetto Java, presente nel server container e totalmente all'oscuro del browser, si possono depositare informazioni legate a una pseudoidentità (come per esempio il fatto che il client si sia identificato e autenticato).

Un esempio di utilizzo dell'oggetto Session si può vedere nel servlet fornito dal professore chiamato "SessionCounter".

Possibile domanda da orale: c'è bisogno del login per personalizzare una pagina? no, basta il concetto di session tracking. A cosa serve il login? Il login serve ad associare ad una pseudoidentità un'identità autenticata allo scopo di fare autorizzazioni.

In lezioni > 8-Servlet-base > progetto eclipse-esempi-base-servlet e servlet con JDBC, ci sono progetti con cui possiamo lavorare per scavare più a fondo in questi argomenti.

Prerequisiti

Connessione a database con JDBC

Informazioni

Informazioni sul progetto.

Nuovo progetto caricato su beep: "Esercizio gestione spese di trasferta", è commentato (slide e readme.md) e contiene una base di dati. Contiene concetti molto importanti su JDBC.

8 JDBC

Gli argomenti di questa lezione verranno corredati alla visualizzazione di un progetto nella cartella su Beep delle lezioni > Servlet-base > Progetto eclipse servlet con JDBC.

JDBC è un architettura molto elegante per il mondo Java e serve per mettere in contatto applicazioni Java con DB.

JDBC eredita la sua logica da ODBC, un architettura sviluppata originariamente da Microsoft.

Il vantaggi di JDBC sono legati al mascheramento delle differenze che sono presenti nelle modalità con cui un'applicazione interagisce con una base di dati. Il principale aspetto di JDBC è quello di migliorare la portabilità dei sistemi che fanno uso di DB. Per ottenere questo sfrutta due tecniche: rimuove la necessità che il programmatore conosca le tecniche con cui avviene la connessione coi DB, maschera alcune differenze "dialettali" nell'uso di primitive SQL che alcune basi di dati offrono e altre magari offrono in maniera diversa.

L'architettura JDBC struttura la connessione con la base di dati in maniera intelligente, che consente all'applicazione di ignorare l'ospetifico modello della base di dati. Il modello utilizzato è **stratificato**: l'**applicazione Java** vede solo un'**API** (application programming interface) al di sotto del quale c'è un ambiente di runtime che, a fronte di un particolare modello di base di dati, carica un determinato Driver. Per un sistema operativo, un Driver è un programma speciale che contiene le istruzioni di gestione di una determinata periferica, questo stesso meccanismo è stato replicato per sviluppare JDBC. Più in dettaglio, l'API utilizza un servizio chiamato **Driver Manager** che ha lo scopo di caricare un determinato **Driver** a seconda di con quale base di dati dovrà interagire.

L'effettiva comodità di JDBC risiede nei vantaggi che il programmatore ne trae, ovvero che dovrà interagire con una serie di oggetti di utilità: Connection, Statement, ResultSet, SQLException (Driver e DriverManager li vedremo una sola volta). L'intera architettura JDBC è fatta da queste classi.

JDBC è utilizzato anche al di fuori dell'ambiente web.

Per l'uso di JDBC all'interno di una Servlet si usa il seguente workflow (pattern):

- connessione;
- preparazione ed esecuzione di query;
- precessamento dei risultati;
- disconnessione (molto importante, siccome la base di dati è una risorsa comune a tutte le servlet, lasciarla "libera" è essenziale, talvolta la disconnessione precede anche il processamento dei risultati; un metodo più avanzato per distribuire in maniera efficiente le connessioni è il cosiddetto "connection pooling", più informazioni sul video di questo argomento).

Nel momento il programmatore non ha controllo sul ciclo di vita della propria applicazione (come nel caso dei Servlet) si prevedono dei pattern specifici riguardanti le quattro operazioni appena elencate.

Connessione: avviene nel momento di creazione della Servlet, cioè nella funzione init() della Servlet

(nelle slide della lezioni ci sono più informazioni sull'esatto procedimento), i parametri (url, utente, pass, modello DB) necessari alla connessione al DB sono tipicamente inseriti all'interno del file xml di configurazione dell'applicazione web;

oss. tutti i blocchi riguardanti l'interazione con DB saranno contenuti in blocchi critici (try... catch... finally), nell'esercizio di gestione spese di trasferta su Beep c'è una variante molto carina, che abbrevia molto la lunghezza del codice.

Disconnessione: simmetricamente la disconnessione è eseguita nel metodo destroy(), per cui la servlet ha una connessione disponibile per tutta la sua durata.

La connessione e disconnessione sono le operazioni più costose all'interno di un DB, dunque ci sono tecniche avanzate che permettono di sfruttare al meglio le connessioni (vedi connection pooling).

Questo schema lo vedremo in tutti i progetti caricati.

In questo schema c'è un grande problema. Un processore di query è un componente che riceve una stringa query SQL, lo analizza (parsing), costruisce l'albero sintattico della query, lo ottimizza, lo compila in un piano di accesso disco e infine lo esegue. Il grande problema è che non c'è bisogno di ricompilare il piano di accesso disco per tutte quelle query che sono "simili" fra di loro (cioè che differiscono per alcune variabili), quindi mandare query a ricompilare continuamente spreca un sacco di tempo. Come soluzione a questo problema intervengono le **parametric prepare statement**, che utilizzano degli scheletri di istruzioni con dei "?" al posto dei valori che potrebbero cambiare, che sono dette variabili di binding, e vengono compilati in piani di accesso una sola volta ed eseguiti quante volte si vuole con le variabili che si vuole. Se una query è molto complessa e la sua preparazione è molto onerosa, sfruttare questi prepare può rappresentare un vantaggio molto ingente.

"Mai mai mai fare query al volo, è sempre buona pratica prepararle prima".

Le **SQL injection** sono un problema che riguardando l'esecuzione di query che sono frutto della concatenazione di una parte sicura con una parte insicura.

Per parte insicura si intendono stringhe senza controlli di formattazione da parte del programmatore. Questi problemi insorgono tipicamente quando la query è formata da una parte programmata e una parte che è il risultato di un parametro o un campo inserito direttamente (senza controllo da parte del programmatore) dall'utente. Per esempio se usiamo una query "select * where ID = " + "[input dell'utente]", un utente male intenzionato potrebbe scrivere "1 OR 1=1", ricevendo quindi come risposta l'intera tabella dal database, in quanto 1=1 sempre vero. Il concetto è che dobbiamo evitare di dare all'utente la possibilità di poter modificare l'albero sintattico della query (nel caso appena mostrato è successo quando l'utente ha inserito "OR") Questo meccanismo può creare grandi falle di sicurezza. Il rimedio a questo problema è l'utilizzo dei prepare statement, infatti preparando le query, l'albero sintattico viene precreato e quindi l'utente può solo e soltanto inserire parametri, che se sono mal formati, la base di dati produce un errore, invece di rilevare informazioni che non dovrebbe.