Beaver®

Healthcare Research & Pharmacoepidemiology

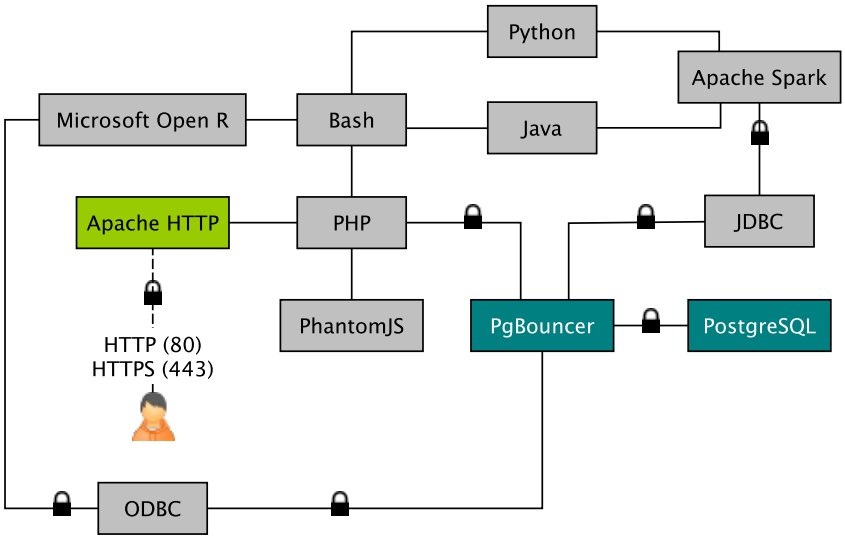
http://www.chrp.it/beaver/
Ultimo aggiornamento: 29/07/2021

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

ARCHITETTURA E DATI

ARCHITETTURA SOFTWARE

Beaver è una piattaforma web per la conduzione automatizzata di studi farmaco-epidemiologici osservazionali su dati clinico-amministrativi, ovvero l'applicazione di protocolli scientifici per la selezione di coorti e l'esecuzione di procedure statistiche parametriche. L'architettura prevede il dialogo tra diversi componenti e la presenza di due ulteriori applicazioni web dedicate ad analisi statistiche personalizzate e al monitoraggio di particolari tipologie di elaborazioni.



Il database utilizzato si chiama PostgreSQL e contiene gli schemi dell'applicazione Beaver e i flussi Regionali normalizzati, organizzati in schemi.

In verde chiaro sono riportati i servizi web a cui ha accesso l'utente:

- Apache HTTP: rappresenta il servizio web principale da cui è possibile autenticarsi e utilizzare Beaver.

In grigio sono riportati gli interpreti e altri software:

- PHP: engine di esecuzione dell'applicazione Beaver.

- Bash: console Bash di sistema, utilizzata per consentire la gestione delle pipeline di esecuzione.
- Java: interprete del linguaggio Java e Scala, utilizzato per l'esecuzione di Apache Spark.
- Python: interprete del linguaggio Python.
- Apache Spark: framework di analisi e calcolo distribuito impiegato per le operazioni ETL e l'esecuzione di particolari algoritmi sui dati.
- Microsoft Open R: versione ottimizzata e Open-Source di R, un interprete del linguaggio R per l'analisi statistica.
- ODBC: driver di sistema utilizzato per il dialogo tra MRO e PostgreSQL.
- JDBC: driver Java utilizzato da Spark per il dialogo con PostgreSQL.
- PhantomJS: browser web programmabile e senza interfaccia grafica, usato per convertire i report HTML in formato PDF.
- PgBouncer: pooler di connessioni al database. Serve a limitare le risorse della macchina dedicate al database, garantendo al tempo stesso un numero di connessioni simultanee accettabile.

AUTENTICAZIONE

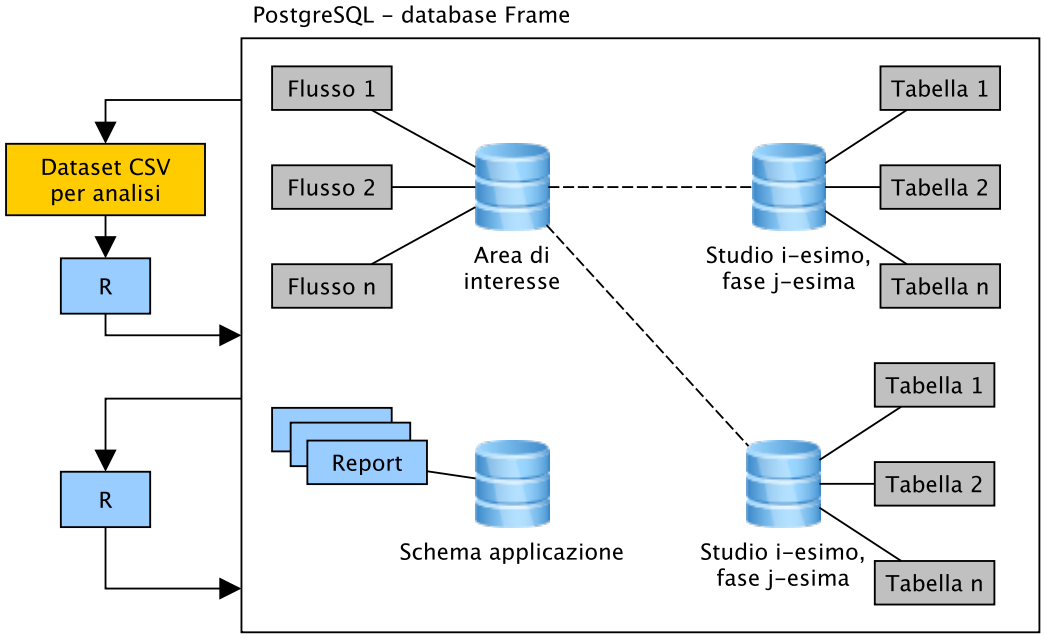
Beaver implementa un sistema di autenticazione interno che non è in alcun modo legato a servizi di dominio come Active Directory. Dal pannello di amministrazione di Beaver è possibile gestire le utenze (creazione, modifica, eliminazione).

L'accesso al database PostgreSQL e ai risultati generati avviene sempre tramite autenticazione. Più in generale, qualsiasi tentativo di accesso prevede una fase di autenticazione, sia che venga effettuato dall'utente (anche a livello di chiamate web asincrone), sia che avvenga tra le componenti interne dell'architettura.

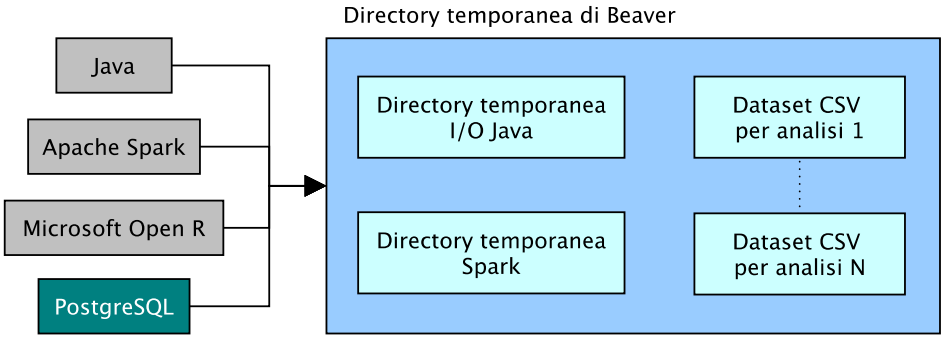
DATI E RISULTATI

I dati sono costituiti dai flussi Regionali. Durante le procedure ETL, i file CSV forniti vengono opportunamente standardizzati, partizionati e indicizzati all'interno di tabelle organizzate in schemi. Ciascuno schema rappresenta un'area di interesse, ovvero un insieme di flussi distinti contenenti eventi associati a soggetti che presentano un particolare profilo clinico. Le aree di interesse costituiscono le popolazioni da cui estrarre le coorti di interesse. Le procedure ETL, infatti, vengono eseguite *una-tantum* per ciascuna area di interesse finché non risulti necessario ricorrere a una versione aggiornata dei flussi.

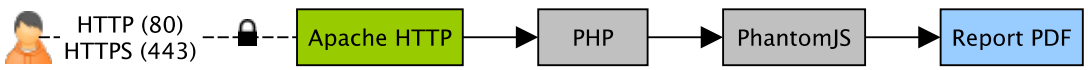
I flussi vengono utilizzati in sola lettura per operazioni di record-linkage automatizzate, generate da Beaver sulla base delle istruzioni fornite dall'utente. I protocolli di studio prevedono l'esecuzione di diverse fasi sequenziali, ciascuna delle quali consente di procedere all'estrazione e alla generazione di tutte le informazioni utili agli obiettivi dello studio. Al termine di ogni fase vengono create delle tabelle organizzate in determinati schemi all'interno del database. I report generati da R sulla base dei dati estratti al termine delle diverse fasi vengono archiviati in formato JSON nel database. L'ultima fase dello studio permette all'utente di eseguire delle analisi parametriche (scelta e definizione di modelli statistici) e, analogamente a quanto fatto per la reportistica delle fasi precedenti, memorizza i risultati aggregati in formato JSON all'interno del database.



La stima dei modelli statistici avviene per mezzo dell'esportazione di alcuni dataset CSV da parte di PostgreSQL, i quali vengono successivamente letti e caricati in memoria da R. Il file system della macchina in cui è installato Beaver deve disporre di una cartella temporanea predisposta all'archiviazione di questi files e di file temporanei creati da Java durante l'esecuzione di Apache Spark.



La produzione del report finale avviene per mezzo di PhantomJS, un browser web batch programmabile che tramite uno script JS renderizza in formato PDF le pagine HTML contenenti grafici e tabelle che l'utente desidera esportare.



Il report PDF attualmente può essere scaricato dall'utente tramite il proprio browser web e ne viene creata una copia sulla macchina in cui è installato Beaver.

AGGIORNAMENTO DI BEAVER

I file che compongono il software Beaver, ovvero un insieme di cartelle, script e altro, sono gestiti tramite il software di versioning Git. Lo scopo di Git è garantire la distribuzione e l'aggiornamento dei file nei diversi siti di installazione. Per usufruire di questa funzionalità è necessario che la macchina su cui è installato Beaver abbia accesso al server Git gestito dall'Università degli Studi di Milano-Bicocca. L'operazione di aggiornamento del codice avviene tramite l'esecuzione di un comando di sistema che l'amministratore della macchina esegue da terminale, previa autenticazione con chiave pubblica al server Git.

INDICATORI MINISTERIALI

Beaver è fruibile in due versioni. Beaver *Full* offre all'utente avanzato tutti gli strumenti necessari a implementare il protocollo di uno studio di coorte generico tramite delle procedure guidate. Beaver *Light* automatizza i protocolli descritti nel manuale del Ministero per il calcolo degli indicatori PDTA. In questa seconda versione, l'utente ha accesso a una schermata molto semplice da cui ha la possibilità di selezionare l'area di interesse, l'anno di riferimento e la variabile di stratificazione (usata per il calcolo degli indicatori stratificati). L'elaborazione può essere monitorata in tempo reale e, una volta terminata, l'utente ha accesso alla reportistica.

REQUISITI DI SISTEMA

Di seguito sono elencati i requisiti di sistema previsti da Beaver. Un dimensionamento bilanciato deve necessariamente tenere conto della dimensione del dato, o in altri termini della dimensione della popolazione residente nella Regione presso cui deve essere installata la piattaforma. I valori riportati pertanto sono puramente indicativi.

Elemento	Minimo	Consigliato
Sistema Operativo	Ubuntu Server 18.04.5 LTS, oppure Red Hat Enterprise 8	Ubuntu Server 20.04.2 LTS
(1) CPU	da 4 a 8	da 16 a 32
(2) RAM	da 16 GB a 32 Gb	da 64 GB a 128 GB
(3) Spazio Database	2 volte il dato originale	3 volte il dato originale
Spazio Applicazione e pacchetti di sistema	16 GB	32 GB
Configurazione RAID (se previsto)	1+0	1+0
Filesystem	EXT4	XFS

(1) Il numero di CPU determina il livello di parallelismo e la velocità di esecuzione di alcune operazioni.

(2) La RAM dipende dalla dimensione del dato originale e dal numero di CPU.

(3) Lo spazio richiesto su disco dal Database dipende dalla dimensione del dato originale.

CONFIGURAZIONE DI RETE

È richiesta l'apertura di alcune porte di rete:

Servizio	Porta	Protocollo	Destinazione
Server web	80, oppure 443	HTTP, oppure HTTPS	Intranet
Git (servizio di versioning)	22 e 9418	SSH e GIT	Server Git ospitato presso Università degli Studi di Milano-Bicocca
Console SSH remota per amministrazione	22	SSH	Intranet
Notifiche email (opzionale)	25	SMTP	Internet
(4) Installazione/aggiornamento pacchetti	80 e 443	HTTP e HTTPS	Internet

(4) Per l'installazione e l'aggiornamento di alcuni pacchetti di sistema, come ad esempio le librerie utilizzate da R, è preferibile (e in alcuni casi potrebbe essere necessario) consentire l'accesso a Internet. Una volta terminata l'installazione o l'aggiornamento è possibile ripristinare regole firewall più restrittive. Per l'installazione delle librerie R, è necessario che il sistema abbia accesso ai seguenti URL: https://mran.revolutionanalytics.com, https://cran.revolutionanalytics.com

APPENDICE A – PACCHETTI RICHIESTI (UBUNTU SERVER 18.04.5 LTS)

Di seguito è riportato l'elenco dei principali pacchetti richiesti per il corretto funzionamento di Beaver su sistemi [1] Ubuntu Server 18.04.5 LTS:

Nome pacchetto	Note
`postgresql-12` `postgresql-client-12`	PostgreSQL >= 12.x (https://www.postgresql.org/download/linux/ubuntu/)
`pgbouncer`	PgBouncer >= 1.9 (https://pgbouncer.github.io/). Ppotrebbe essere necessario installare anche `libevent` e `libevent-devel`
`Apache2`	Apache 2.x
`python`	Python >= 3.6

`php7.3`
`php7.3-common`
`php7.3-mbstring`
`php7.3-pgsql`
`php7.3-xml`
`libapache2-mod-php7.3`
	PHP >= 7.3

Java JRE Server 8	Java JRE Server 8 >= 1.8.0_251 https://www.oracle.com/it/java/technologies/javase-server-jre8-downloads.html
`odbc-postgresql`	Driver ODBC
Microsoft R Open (MRO)	>= 4.x (https://mran.microsoft.com/download/)

`gcc`
`make`
`gfortran`
`g++`
`libiodbc2-dev`
`libssl-dev`
`libcairo2-dev`
`libxml2-dev`
`unixodbc-dev`
`libudunits2-dev`
`libgeos-dev`
`libgdal-dev`
	Pacchetti necessari alla compilazione delle librerie R (l'elenco potrebbe subire variazioni nel tempo)

`git`	
`fontconfig` `ttf-mscorefonts-installer`	Pacchetti necessari alla generazione corretta dei report PDF generati da Beaver (l'elenco potrebbe subire variazioni nel tempo)
`monaco-font`	https://github.com/cstrap/monaco-font
`phantomjs-2.1.1-linux-x86_64`	http://phantomjs.org/

(1) È previsto il supporto anche alla versione Ubuntu Server 20.04.2 LTS

APPENDICE B – PACCHETTI RICHIESTI (RED HAT ENTERPRISE 8)

Di seguito è riportato l'elenco dei principali pacchetti richiesti per il corretto funzionamento di Beaver su sistemi Red Hat Enterprise 8:

Nome pacchetto	Note

`xorg-x11-server-common`
`xorg-x11-server-utils`
`firefox`
	X11 e Firefox per debug post installazione (richiesto nel caso in cui non sia possibile collegarsi tramite HTTPS alla macchina da remoto)

`postgresql12`
`postgresql12-contrib`
`postgresql12-odbc`
	PostgreSQL >= 12.x (https://www.postgresql.org/download/linux/redhat/)

`pgbouncer`	PgBouncer >= 1.9 (https://pgbouncer.github.io/). Ppotrebbe essere necessario installare anche `libevent` e `libevent-devel`
`python`	Python >= 3.6
`httpd`	Apache 2.x

`php73w`
`php73w-opcache`
`php73w-pecl-apcu`
`php73w-cli`
`php73w-pear`
`php73w-pdo`
`php73w-pgsql`
`php73w-pecl-memcache`
`php73w-pecl-memcached`
`php73w-gd`
`php73w-mcrypt`
`php73w-xml`
`php73w-mbstring`
	PHP >= 7.3

Java JRE Server 8	Java JRE Server 8 >= 1.8.0_251 https://www.oracle.com/it/java/technologies/javase-server-jre8-downloads.html
`unixODBC-devel`	Driver ODBC
Microsoft R Open (MRO)	>= 4.x (https://mran.microsoft.com/download/)

`gcc`
`make`
`gfortran`
`g++`
`libiodbc-dev`
`openssl-devel`
`cairo-devel`
`libxml2-devel`
	Pacchetti necessari alla compilazione delle librerie R (l'elenco potrebbe subire variazioni nel tempo)

`git`	

`curl`
`cabextract`
`fontconfig`
`xorg-x11-font-utils`
`msttcore-fonts-installer`
`mkfontscale`
`mkfontdir`
	Pacchetti necessari alla generazione corretta dei report PDF generati da Beaver (l'elenco potrebbe subire variazioni nel tempo)

`monaco-font`	https://github.com/cstrap/monaco-font
`phantomjs-2.1.1-linux-x86_64`	http://phantomjs.org/