



UNIVERSITÀ DEGLI STUDI DI MILANO
FACOLTÀ DI SCIENZE E TECNOLOGIE

Corso di Laurea in Informatica Musicale

MOONCLOUD RECOMMENDATION SYSTEM

Relatore:

Claudio Agostino Ardagna

Correlatore:

Valerio Bellandi

Tesi di Laurea di:

Andrea Michele Albonico

Matricola: 886667

Anno Accademico 2018/2019

Ringraziamenti

Andrea Michele Albonico

Prefazione

I sistemi di raccomandazione (*Recommendation System*) hanno avuto un forte sviluppo negli ultimi decenni e nascono proprio con lo scopo di identificare quegli oggetti (detti generalmente *item*) all'interno di un vasto mondo di informazioni che possono essere di nostro interesse e tanto maggiore è il grado di conoscenza dell'individuo e tanto più vengono ritenuti affidabili.

Il motivo di questo successo risiede nella riuscita integrazione di tali sistemi in applicazioni commerciali, soprattutto nel mondo dell'E-commerce e nel fatto che sono in grado di aiutare un utente a prendere una decisione, che sia la scelta di un film per l'uscita con gli amici il sabato sera, di una playlist da ascoltare durante un viaggio in auto o in un momento di lettura, e via discorrendo.

Moon Cloud è una piattaforma erogata come servizio che fornisce un meccanismo di *Security Governance* centralizzato. Garantisce il controllo della sicurezza informatica in modo semplice e intuitivo, attraverso attività di test e monitoraggio periodiche e programmate (*Security Assurance*). L'obiettivo di questa tesi è stato quello di aggiungere, al già presente sistema per la scelta dei controlli all'interno delle attività di test, un sistema di raccomandazione che possa consigliare all'utente delle possibili *Evaluation* rispetto al target indicato; in questo modo anche l'utente meno esperto può usufruire dei servizi offerti da Moon Cloud in modo semplice e intuitivo.

La tesi è organizzata come segue:

Capitolo 1 – Introduzione a Moon Cloud in questo capitolo viene descritta la piattaforma Moon Cloud e il suo funzionamento in ambito di Security Assurance.

Capitolo 2 – Tecnologie utilizzate in questo capitolo vengono presentati gli studi e le analisi di soluzioni esistenti, studi delle tecnologie utilizzate per la realizzazione del progetto.

Capitolo 3 – Collaborative filtering in questo capitolo viene descritto in modo più approfondito gli studi compiuti sui Filtri Collaborativi

che hanno portato alla realizzazione dei sistemi di raccomandazione proposti nella soluzione implementata per la piattaforma Moon Cloud, inoltre verranno mostrate le relative porzioni di codice.

Capitolo 4 – Descrizione della soluzione in questo capitolo viene descritta in maniera dettagliata la realizzazione dell'applicativo, analizzando quali sono state le difficoltà maggiori, i risultati ottenuti e l'uso che se ne è fatto.

Capitolo 5 – Conclusioni in questo capitolo vengono esposte le conclusioni e i possibili sviluppi futuri delle attività svolte e del sistema realizzato.

Indice

Prefazione	v
1 Introduzione a Moon Cloud	1
1.1 Moon Cloud overview	1
1.2 Processo di Evaluation	5
2 Tecnologie utilizzate	7
2.1 Perché Python e Django	7
2.2 Docker	11
2.3 Strutture dati gerarchiche	12
2.3.1 The Adjacency List Model	13
2.3.2 The Nested Set Model	14
2.4 Sistemi di raccomandazione	17
2.4.1 Content-based filtering	19
2.4.2 Collaborative filtering	19
2.4.3 Cold Start problem	21
3 Collaborative filtering	23
3.1 Memory-based	23
3.1.1 User-based filtering	23
3.1.2 Item-based filtering	24
3.1.3 Hybrid filtering	25
4 Descrizione della soluzione	27
5 Conclusioni	31
Bibliografia	33

Elenco delle figure

1.1	Security Compliance Evaluation	4
2.1	Schema generico di funzionamento di un applicativo web sviluppato con Django	10
2.2	Schematizzazione del contenuto di un Container in Docker . . .	11
2.3	Esempio della rappresentazione gerarchica parziale dei dati nel progetto in questione	12
2.4	Esempio della gestione di dati in modo gerarchico secondo il Nested Set Model, utilizzando dati presi dal database del progetto in questione	14
2.5	Esempio della gestione di dati in modo gerarchico secondo il Nested Set Model, utilizzando dati presi dal database del progetto in questione	15
2.6	Categorizzazione generale dei sistemi di raccomandazione . . .	17
3.1	Esempio di applicazione di un sistema di raccomandazione User-based	24
3.2	Esempio di applicazione di un sistema di raccomandazione Item-based	25
4.1	Struttura dell'applicativo sfruttando il framework proposto da Django	28
4.2	Struttura del database	29

Elenco delle tabelle

2.1	Esempio di una possibile tabella per gestire dati in modo gerarchico secondo l'Adjacency List Model	13
2.2	Esempio di una tabella per gestire dati in modo gerarchico secondo il Nested Set Model	15

Capitolo 1

Introduzione a Moon Cloud

In questo capitolo verrà descritto in modo più approfondito il funzionamento della piattaforma Moon Cloud e unitamente al motivo dell'implementazione della soluzione proposta.

1.1 Moon Cloud overview

La diffusione di sistemi *Information and Communications Technology* (acronimo ICT) ha avuto luogo nella maggiorparte degli ambienti lavorativi e privati in termini di servizi offerti, automazione di processi e incremento delle performance. L'uso di questa tecnologia ha assunto importanza a partire dagli anni novanta come effetto del boom di Internet e al giorno d'oggi le professionalità legate al mondo dell'ICT crescono in numero e si evolvono per specificità, per operare in ambienti fortemente eterogenei ma sempre più interconnessi fra di loro come il Cloud Computing, i Social Newtwork, il Marketing Digitale, i Sistemi IoT, la Realtà Virtuale, etc.

Il Cloud Computing ha portato un rivoluzionario paradigma nella creazione di un nuovo business virtuale accessibile in qualunque momento e in qualunque luogo; esso sfrutta le tecnologie messe a disposizione dai sistemi ICT come le operazioni di virtualized computing, internet e distributed computing, provvedendo un sistema integrato molto potente. Google, Microsoft, Amazon sono un esempio di aziende che forniscono servizi di Cloud Computing in business ICT. Si può definire il Cloud Computing come l'abilità di accedere a risorse (come database o applicazioni) in tutto il mondo attraverso una rete in poco tempo.

Gli immensi benefici del Cloud in termini di flessibilità, consumo delle risorse e gestione semplificata, la rende la prima scelta per utenti e industrie per il deploy dei loro sistemi IT. Tuttavia il Cloud Computing solleva diverse

problematiche legate alla mancanza di fiducia e trasparenza dove i clienti necessitano di avere delle garanzie sui servizi Cloud ai quali si affidano; spesso i fornitori di questi servizi non forniscono ai clienti le specifiche riguardanti le misure di sicurezza messe in atto.

Negli ultimi anni, sono state sviluppate tecniche e modi per rendere sicuri questi sistemi e proteggere i dati degli utenti, portando alla diffusione di approcci eterogenei che incrementano la confusione negli utenti. Tecniche tradizionali di verifica della sicurezza basati su metodi di analisi statistica non sono più sufficienti e devono essere integrati con processi di raccolta di prove (in inglese *evidence*) da sistemi Cloud in produzione e funzionanti. In generale il *Cloud Security* definisce i modi, come crittazione e controllo degli accessi, per proteggere attivamente gli asset da minacce interne ed esterne, e fornire un ambiente in cui i clienti possano affidarsi e interagire in totale sicurezza. Tutto questo non basta a rendere il Cloud degno di fiducia e trasparente, per questo sono state introdotte tecniche di *Security Assurance*, delle garanzie che permettono di ottenere la fiducia necessaria nelle infrastrutture e/o nelle applicazioni di dimostrare il rispetto di certe proprietà di sicurezza, e che operino normalmente anche se subiscono attacchi; grazie alla raccolta e allo studio di *evidence* è possibile che venga accertata la validità e l'efficienza delle proprietà di sicurezza messe in atto.

Il prezzo che si paga per i benefici di questa tecnologia è dato dall'incremento di violazioni di sicurezza, che oggi preoccupa tutte le aziende e di conseguenza anche i loro clienti, con l'incremento del rischio di fallimento per i servizi più importanti dovuti a violazioni della privacy e al furto di dati.

Il mercato sta lentamente notando che non è l'inadeguamento tecnologico dei sistemi di sicurezza che incrementa il rischio delle violazioni di sicurezza; piuttosto, la mal configurazione e l'errata integrazione di questi sistemi nei processi di business. [2] L'utilizzo di sistemi di sicurezza e di controllo migliori non garantisce in modo assoluto la sicurezza dell'infrastruttura; per garantire ciò è necessario implementare un processo continuo di diagnostica e verifica della corretta configurazione dei controlli, supervisionando il loro comportamento, accertandosi che sia quello aspettato.

Il *Security Assessment* diventa allora un aspetto importante specialmente negli ambienti Cloud e IoT. Questo processo, costituito da un insieme di attività mirate alla valutazione del rischio in sistemi IT, deve essere portato avanti in modo continuo e olistico, per correlare le *evidence* raccolte da sempre maggiori meccanismi di protezione. [1]

Moon Cloud è una soluzione PaaS (acronimo inglese di Platform as a Service) che fornisce una piattaforma B2B (Business To Business) innovativa per verifiche, diagnostiche e monitoraggio dell'adeguatezza dei sistemi ICT rispetto

alle politiche di sicurezza, in modo continuo e su larga scala. Moon Cloud supporta una semplice ed efficiente *ICT Security Governance*, dove le politiche di sicurezza possono essere definite dalle compagnie stesse (a partire da un semplice controllo sulle vulnerabilità a linee guida di sicurezza interna), da entità esterne, imposte da standard oppure da regolamentazioni nazionali o internazionali.

La sicurezza di un sistema o di un insieme di asset dipende solo parzialmente dalla forza dei singoli meccanismi di protezione isolati l'uno dall'altro; infatti, dipende anche dall'abilità di questi meccanismi di lavorare continuamente in sinergia per provvedere una protezione olistica. In più, quando i sistemi Cloud e i servizi IoT sono coinvolti, le dinamiche di questi servizi e la loro rapida evoluzione rende il controllo dei processi all'interno dell'azienda e le politiche di sicurezza più complesse e prone ad errori.

I requisiti ad alto livello fondamentali per poter garantire le Security Assurance sono:

sistema olistico è richiesta una visione globale e pulita dello status dei sistemi di sicurezza; inoltre è cruciale distribuire lo sforzo degli specialisti in sicurezza per migliorare il processo e le politiche messe in atto. Si parte da delle valutazioni fatte manualmente a quella semi-automatiche che vengono usate per ispezionare i meccanismi di sicurezza.

monitoraggio continuo ed efficiente è necessario un controllo continuo che valuti l'efficienza dei sistemi di sicurezza per ridurre l'impatto dell'errore umano, soprattutto dal punto di vista organizzativo. La coesistenza di componenti in conflitto o la mancata configurazione dovuta al cambiamento dell'ambiente possono essere scenari che richiedono un monitoraggio e un aggiornamento continuo.

singolo punto di management avere un solo punto d'accesso in cui poter gestire tutti gli aspetti relativi alla sicurezza, permette di avere sotto controllo le politiche di sicurezza. Inoltre disporre di un inventario degli asset da proteggere permette di poter conoscere quali meccanismi di protezioni applicare.

reazioni rapide a incidenti di sicurezza spesso la reazione ad incidenti di sicurezza è ritardata da due fattori: il tempo richiesto per rilevare l'incidente e il tempo per analizzare il motivo dell'accaduto; e avere un sistema che implementa un monitoraggio continuo permette di venire a conoscenza di questi problemi in breve tempo e agire di conseguenza.

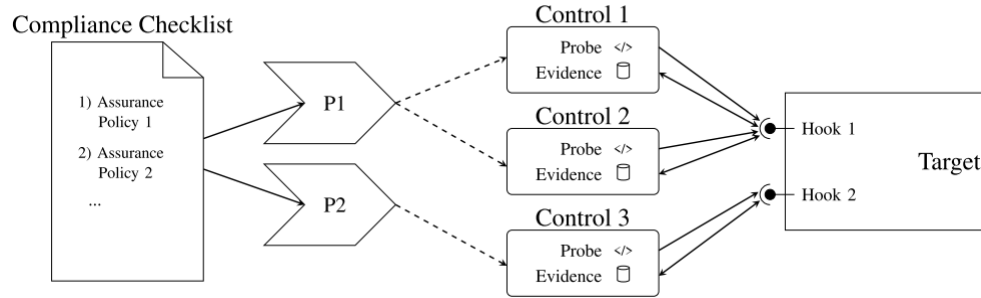


Figura 1.1: Security Compliance Evaluation

Moon Cloud è basato su una tecnica di Security Assurance garantendo che tutte le attività aziendali si compiano seguendo i requisiti prestabiliti da appropriate politiche e procedure precedentemente definite.

Viene definita una *Security Compliance Evaluation*, un processo di verifica a cui un target viene sottoposto e il cui risultato deve soddisfare i requisiti richiesti da standard e politiche. A partire da questi processi di controllo, che devono a loro volta essere affidabili, si ottengono delle evidenze; queste ultime possono essere raccolte monitorando l'attività del target oppure, come già menzionato, sottoponendo il target a scenari critici o di testing. In particolare, una Security Compliance Evaluation è un processo di verifica dell'uniformità di un certo target a una o più politiche attraverso una serie di controlli che a seconda del valore booleano (successo pari a 1 o insuccesso pari a 0) associato ad ogni controllo viene prodotto un valore booleano per le politiche; se un target supera tutti i controlli, quindi tutti i valori booleani associati ad ogni controllo è pari al successo, a cui è sottoposto allora significa che rispetta la politica scelta.

1.2 Processo di Evaluation

Moon Cloud implementa il processo di Security Compliance Evaluation in Figura 1.1 usando controlli di monitoraggio o di test personalizzabili. Inoltre garantisce, oltre a tutti i requisiti ad alto livello elencati precedentemente 1.1, anche i seguenti:

Moon Cloud è una piattaforma Cloud centralizzata presentando una visione olistica dello stato di sicurezza di un dato sistema.

Moon Cloud implementa un sistema di Security Assurance Evidence-based continuo, implementato come processo di Compliance, basato su politiche custom o standard.

Moon Cloud è offerto come un servizio (PaaS), dove le attività di Evaluation possono essere facilmente ed efficientemente configurate su un target asset, senza l'intervento dell'uomo.

Moon Cloud permette di schedulare delle ispezioni automatiche, grazie all'inventario di asset protetto.

Moon Cloud Evaluation Engine può ispezionare dall'interno un sistema, gestendo così delle minacce interne; permettendo anche reazioni rapide a incidenti di sicurezza e veloci rimedi, grazie alla raccolta continua di evidence.

In generale, l'architettura di Moon Cloud è costituita da un'Assurance Manager che gestisce i processi di Evaluation attraverso un set di *Execution Cluster*; ognuno dei quali gestisce ed esegue un set di probe che collezionano le evidence necessarie per effettuare i processi di valutazione. Tutte le attività di collezione sono eseguite dal probe, ognuno dei quali è uno script Python fornito come una singola immagine di Docker, che viene inizializzata quando viene triggerata una Evaluation ed è distrutta quando il processo di Evaluation è terminato.

Accedendo alla piattaforma di Moon Cloud, l'utente può definire le proprie politiche di sicurezza e attività di Evaluation come espressioni booleane di controlli di sicurezza e altre politiche predefinite. Una volta che una politica viene definita, l'utente può decidere quando schedulare l'Evaluation; e nel momento in cui un processo di Evaluation viene inizializzato, tutti i controlli vengono eseguiti e i risultati dell'espressioni booleane vengono memorizzati e restituiti all'utente. A questo punto l'utente può accedere a questi risultati a diversi gradi di precisione: una visione sommaria e generale di tutte le politiche implementate e dello stato generale del sistema di sicurezza, al risultato di una specifica politica oppure alle evidence raccolte per una Evaluation.

Per poter rendere ancora più intuitivo e semplice da utilizzare un sistema di questa importanza, si è pensato di introdurre un sistema che possa raccomandare agli utenti, in base agli asset che vuole proteggere e monitorare, una serie di Evaluation o politiche da applicare in quei casi; questo permette anche a utenti meno esperti di poter configurare in modo rapido ed efficiente dei meccanismi di protezione da minacce.

Capitolo 2

Tecnologie utilizzate

In questo capitolo verranno descritte le attività preliminari per la realizzazione di questo progetto, le tecnologie utilizzate unitamente alle motivazioni legate all'uso di questi sistemi rispetto ad altri. In particolare nel capitolo successivo verranno approfonditi a livello pratico i sistemi di raccomandazione Memory-based i quali sono stati utilizzati per l'implementazione della soluzione.

2.1 Perché Python e Django

Python Python è un linguaggio di programmazione ad alto livello, orientato agli oggetti, adatto, tra gli altri usi, a sviluppare applicazioni distribuite, scripting, computazione numerica e system testing; ideato e rilasciato pubblicamente per la prima volta nel 1991 dal suo creatore Guido van Rossum, programmatore olandese.

Python supporta diversi paradigmi di programmazione, come quello object-oriented (con supporto all'ereditarietà multipla), quello imperativo e quello funzionale, ed offre una tipizzazione dinamica forte. È fornito di una libreria built-in estremamente ricca, che unitamente alla gestione automatica della memoria e a robusti costrutti per la gestione delle eccezioni fa di Python uno dei linguaggi più ricchi e comodi da usare.

Inoltre è anche semplice da usare e imparare. Python, nelle intenzioni di Guido van Rossum, è nato per essere un linguaggio immediatamente intuibile. La sua sintassi è pulita e snella così come i suoi costrutti, decisamente chiari e non ambigui. I blocchi logici vengono costruiti semplicemente allineando le righe allo stesso modo, incrementando la leggibilità e l'uniformità del codice anche se vi lavorano diversi autori.

Un aspetto inusuale del Python è il metodo che usa per delimitare i blocchi di programma, che lo rende unico fra i linguaggi più diffusi.

```
1 # Testing if two strings are equals
2 def test(got, expected):
3     if got == expected:
4         prefix = ' OK '
5     else:
6         prefix = ' X '
7     print('%s got: %s expected: %s' % (prefix, repr(got), repr(
8         expected)))
9
10 def main():
11     print('verbing')
12     test('hail', 'hailing')
13     test('swiming', 'swimmingly')
14     test('do', 'do')
15
16 if __name__ == '__main__':
17     main()
```

Listing 2.1: Esempio di programma scritto in Python

Nei linguaggi derivati dall'ALGOL come Pascal, C e Perl, i blocchi di codice sono indicati con le parentesi oppure con parole chiave (il C ed il Perl usano `{}`; il Pascal usa `begin` ed `end`). In questi linguaggi è solo una convenzione degli sviluppatori il fatto di indentare il codice interno ad un blocco, per metterlo in evidenza rispetto al codice circostante. In Python invece di usare parentesi o parole chiave, usa l'indentazione stessa per indicare i blocchi nidificati in congiunzione col carattere "due punti" (`:`). Si può usare sia una tabulazione, sia un numero arbitrario di spazi, ma lo standard Python è di 4 spazi.

Python è un linguaggio pseudocompilato: un interprete si occupa di analizzare il codice sorgente (semplici file testuali con estensione `.py`) e, se sintatticamente corretto, di eseguirlo e non esiste una fase di compilazione separata (come avviene in C, per esempio) che generi un file eseguibile partendo dal sorgente. [7]

Django Django è un web framework di alto livello basato su Python che permette di sviluppare rapidamente e con tutti i presupposti per un sistema sicuro, un sito web perfettamente mantenibile, Django si occupa della maggiori grane del sviluppo web, così da permetterti di concentrarti sulla scrittura della tua app; inoltre è open-source e ha una comunità attiva, una documentazione completa e semplice da consultare.

Django aiuta a scrivere software con le seguenti caratteristiche [3]:

Versatile : è usato per la creazione di praticamente tutti i tipi di siti web, a partire a sistemi per la gestione di contenuti e wiki, social network e siti di notizie; può lavorare con qualunque client-side framework, gestire contenuti in quasi tutti i formati (inclusi HTML, RSS feeds, JSON, XML, etc). Internamente permette la scelta e l'implementazione di qualsiasi funzionalità (es. molti dei database più popolari, etc.).

Sicuro : aiuta gli sviluppatori a evitare gli errori più comuni in merito alla sicurezza provvedendo un framework costruito per eseguire le operazioni in modo corretto e sicuro. Ad esempio, Django fornisce un modo sicuro per gestire gli account degli utenti e le relative password, evitando errori comuni come inserire informazioni riguardanti la sessione dell'utente nei cookies dove sarebbero vulnerabili (invece i cookie contengono soltanto una chiave, e i valori effettivi sono salvati nel database) o salvare direttamente una password invece di una hash password.

Mantenibile : il codice di Django è scritto seguendo i principi e i pattern che incoraggiano la creazione di codice mantenibile e riusabile. Inoltre particolare, fa uso del principio "Don't Repeat Yourself" (DRY) così da ridurre al minimo le duplicazioni non necessarie, diminuendo la quantità di codice. Django raggruppa parti di codice letto in moduli seguendo le linee guida del pattern Model View Controller (MVC).

Portatile : Django essendo scritto in Python, un linguaggio multi piattaforma, lo rende indipendente dal sistema operativo eseguito sul server, che sia Linux, Windows o Mac OS X. Per di più, Django è ben supportato da molti web hosting provider, che spesso provvedono a specifiche infrastrutture e documentazione per l'hosting di siti web in Django.

Un tradizionale sito web attende delle richieste HTTP dal browser web (o altri client). Quando viene ricevuta una richiesta, di tipo POST o GET, l'applicazione legge le informazioni contenute nel URL e altri possibili dati a seconda del tipo di richiesta. A seconda della richiesta è possibile che vengano letti o scritti dati da un database o altre operazioni che portino al soddisfacimento della richiesta. A quel punto l'applicazione ritorna una risposta al browser web, spesso in modo dinamico, creando una pagina HTML da mostrare in cui è possibile inserire o recuperare dati in placeholder in un template HTML.

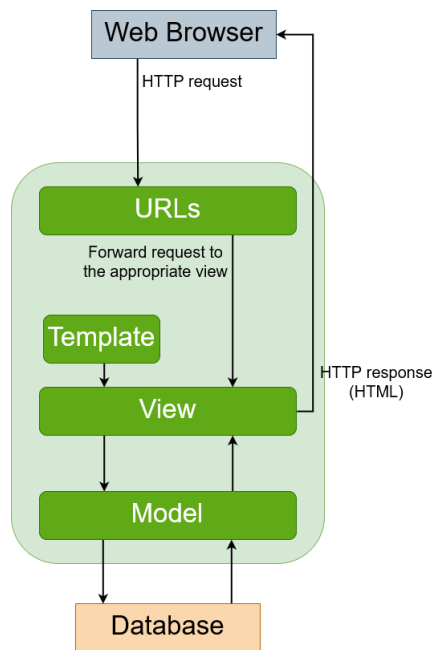


Figura 2.1: Schema generico di funzionamento di un applicativo web sviluppato con Django

Un'applicazione web in Django tipicamente raggruppa il codice che gestisce questi diversi passaggi in file separati [5]:

URL : mentre è possibile processare richieste da qualsiasi URL attraverso una singola funzione, è più mantenibile scrivere diverse funzioni chiamate View per gestire ogni risorsa. un URL mapper è usato per reindirizzare le richieste HTTP alla view corretta in base all'URL della richiesta; inoltre è possibile controllare se nell'URL è presente un particolare pattern di stringhe o numeri, e passare di conseguenza la richiesta alla funzione appropriata come dati da elaborare.

View : una View è una funzione che gestisce le richieste HTTP, e restituisce una risposta HTTP. Le view accedono ai dati necessari per soddisfare la richiesta attraverso i Model, e delegano la formattazione delle risposte ai template.

Model : i Model sono oggetti in Python che definiscono la struttura dei dati dell'applicazione, e provvedono meccanismi per gestirla (add, modify, delete) e query per interpellare il database.

Template : un template è un file di testo che definisce la struttura o il layout di un file (come una pagina HTML), attraverso placeholder per rappre-

sentare il contenuto effettivo. Una View può creare dinamicamente una pagina HTML usando un template HTML, popolandolo con dati presi dal Model.

2.2 Docker

Docker è una piattaforma software che permette di creare, testare e distribuire applicazioni con la massima rapidità. Docker raccoglie il software in unità standardizzate chiamate *container* che offrono tutto il necessario per la loro corretta esecuzione, incluse librerie, strumenti di sistema, codice e runtime. Con Docker, è possibile distribuire e ricalibrare le risorse per un'applicazione in qualsiasi ambiente, tenendo sempre sotto controllo il codice eseguito.

La tecnologia Docker utilizza solitamente il kernel di Linux e le sue funzionalità, come Cgroups e namespace, per isolare i processi in modo da poterli eseguire in maniera indipendente. Questa indipendenza è l'obiettivo dei container: la capacità di eseguire più processi e applicazioni in modo separato per sfruttare al meglio l'infrastruttura esistente pur conservando il livello di sicurezza che sarebbe garantito dalla presenza di sistemi separati.

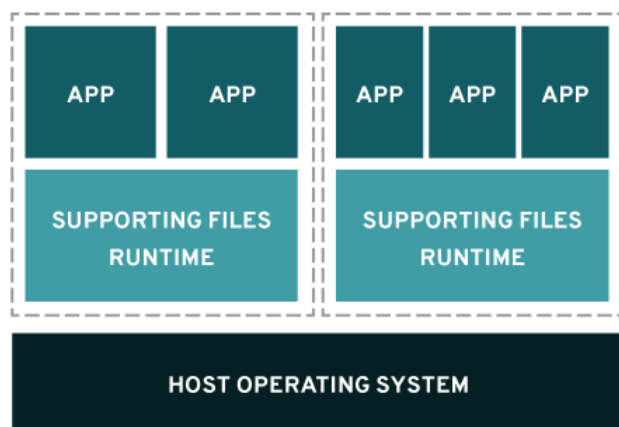


Figura 2.2: Schematizzazione del contenuto di un Container in Docker

Gli strumenti per la creazione di container, come Docker, consentono il deployment a partire da un'*immagine*, ciò semplifica la condivisione di un'applicazione o di un insieme di servizi, con tutte le loro dipendenze, nei vari ambienti. Docker, considera i container come macchine virtuali modulari estremamente leggere, offrendo la flessibilità di creare, distribuire, copiare e spostare i container da un ambiente all'altro, ottimizzando così le app per il cloud.

I contenitori forniscono una modalità standard per impacchettare il codice della tua applicazione, le configurazioni e le dipendenze, in un oggetto singolo e condividono un sistema operativo installato sul server, operando come processi con risorse isolate, assicurando velocità, affidabilità e distribuzioni coerenti, indipendentemente dall'ambiente.

2.3 Strutture dati gerarchiche

Le tabelle di un database relazionale non sono gerarchiche (come nel XML), ma sono delle semplici liste piatte. I dati gerarchici sono costituiti da relazioni padre-figlio che non possono essere rappresentate in modo naturale nelle tabelle di questo tipo. In questo caso, i dati gerarchici sono una collezione di informazioni dove ogni item ha un solo padre e nessuno o più figli (ad eccezione del nodo radice che non ha un nodo padre); questo genere di rappresentazione delle informazioni, simili a un albero, può essere trovato in diversi ambiti di applicazione di un database, incluse discussioni su forum e mailing list, grafici di organizzazione di un business, categorie per gestire contenuti e categorie di prodotti.

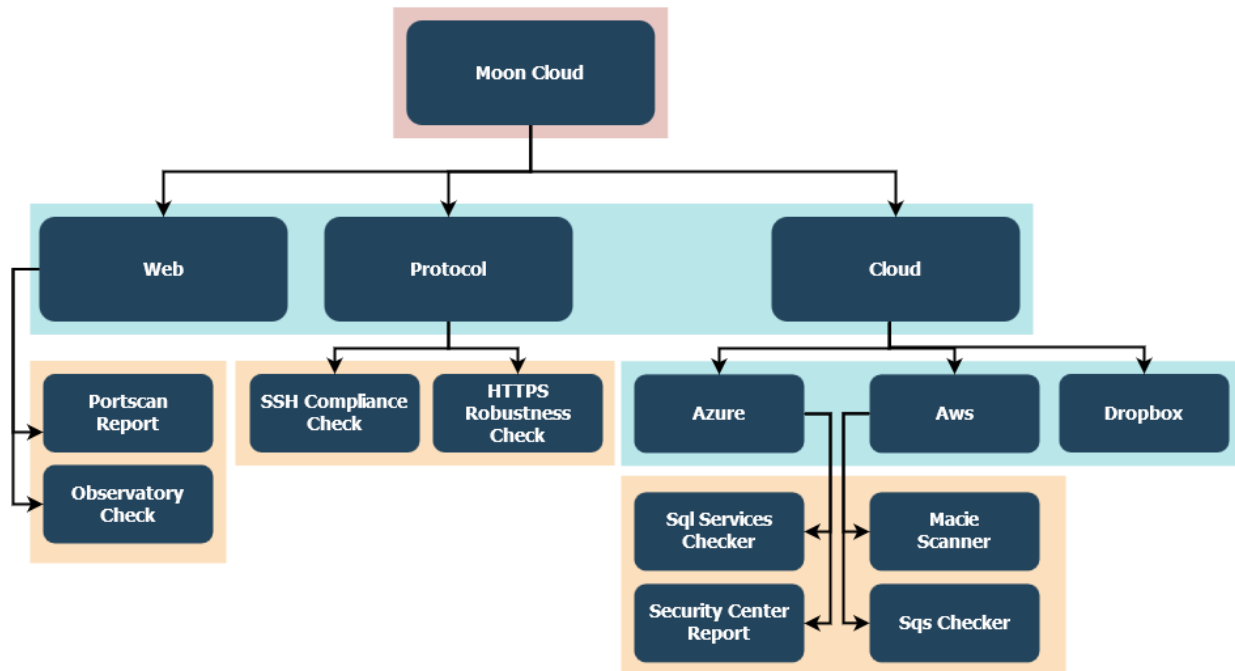


Figura 2.3: Esempio della rappresentazione gerarchica parziale dei dati nel progetto in questione

Ci sono differenti modelli per poter gestire dati in modo gerarchico, i più importanti presi in considerazione sono i seguenti:

2.3.1 The Adjacency List Model

Il primo approccio, e quello di più semplice implementazione, qui descritto è chiamato *Adjacency List Model* o metodo ricorsivo; è definito tale perchè il suo funzionamento si basa su una funzione che itera per tutto l'albero. In questo modello, ogni item (nodo dell'albero) nella tabella contiene un puntatore al suo item padre; invece il nodo radice avrà un puntatore a un valore NULL per l'item padre.

La Tabella 2.1 è un esempio di possibile rappresentazione parziale dei dati nel database implementato in questo progetto secondo questo approccio, seguendo come riferimento la Figura 2.3.

id	name	parent
1	Moon Cloud	NULL
2	Web	1
3	Protocol	1
4	Cloud	1
5	Portscan Report	2
6	Observatory Check	2
7	SSH Compliance Check	3
8	HTTPS Robustness Check	3
9	Azure	4
10	Aws	4
11	Dropbox	4
12	Sql Services Checker	9
13	Security Center Report	9
14	Macie Scanner	10
15	Sqs Checker	10

Tabella 2.1: Esempio di una possibile tabella per gestire dati in modo gerarchico secondo l'Adjacency List Model

Il vantaggio di usare questo modello sta nella sua semplicità di costruzione, soprattutto a livello di codice client-side, e di restituzione dei figli di un nodo. Mentre diventa problematico se si lavora in puro codice SQL e nella maggior parte dei linguaggi di programmazione, è lento e poco efficiente,

perchè è necessaria una query per ogni nodo dell'albero, e visto che ogni query impiega un certo periodo di tempo, questo rende la funzione molto lente quando si lavora con alberi di grandi dimensioni, inoltre molti linguaggi non sono ottimizzati per funzioni ricorsive. Per ogni nodo, la funzione crea una nuova istanza di se stessa e ogni istanza occupa una porzione di memoria e impiega un certo tempo per inicializzarsi, più grande è l'albero e più questo processo sarà portato a termine in maggior tempo.

2.3.2 The Nested Set Model

Il secondo approccio analizzato è il *Nested Set Model*, che permette di osservare la gerarchia in un modo diverso, non come nodi e linee (come se fosse un albero), ma come container innestati.

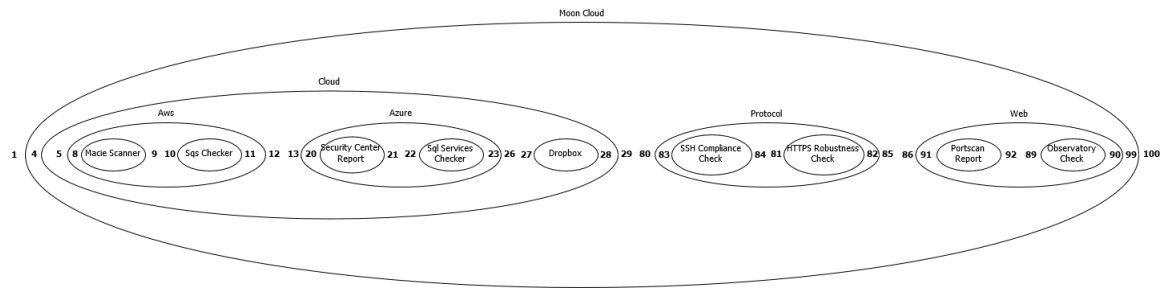


Figura 2.4: Esempio della gestione di dati in modo gerarchico secondo il Nested Set Model, utilizzando dati presi dal database del progetto in questione

Con questo sistema la gerarchia viene mantenuta, secondo il principio cui un nodo padre contiene e suoi figli e questa forma di gerarchia viene mantenuta in tabella attraverso l'uso di due attributi aggiuntivi come è possibile osservare dalla Tabella 2.2 seguente.

Come è possibile osservare dalla Tabella 2.2 la gerarchia dei dati viene rappresentata attraverso l'uso degli attributi 'left' e 'right' per rappresentare l'annidamento dei nodi (il nome delle colonne: 'left' e 'right', hanno significati speciali in SQL; per questo motivo si identificano questi campi con i nomi 'lft' e 'rft').

L'assegnazione di questi valori ad ogni nodo viene effettuata seguendo questo procedimento: ogni nodo dell'albero viene visitato due volte, assegnando i valori in ordine di visita, e in entrambe le visite. Quindi vengono associati ad ogni nodo due numeri, memorizzati come due attributi. Più precisamente

id	name	lft	rght
1	Moon Cloud	1	100
2	Web	86	99
3	Protocol	80	85
4	Cloud	4	29
5	Portscan Report	91	92
6	Observatory Check	89	90
7	SSH Compliance Check	83	84
8	HTTPS Robustness Check	81	82
9	Azure	13	26
10	Aws	5	12
11	Dropbox	27	28
12	Sql Services Checker	22	23
13	Security Center Report	20	21
14	Macie Scanner	8	9
15	Sqs Checker	10	11

Tabella 2.2: Esempio di una tabella per gestire dati in modo gerarchico secondo il Nested Set Model

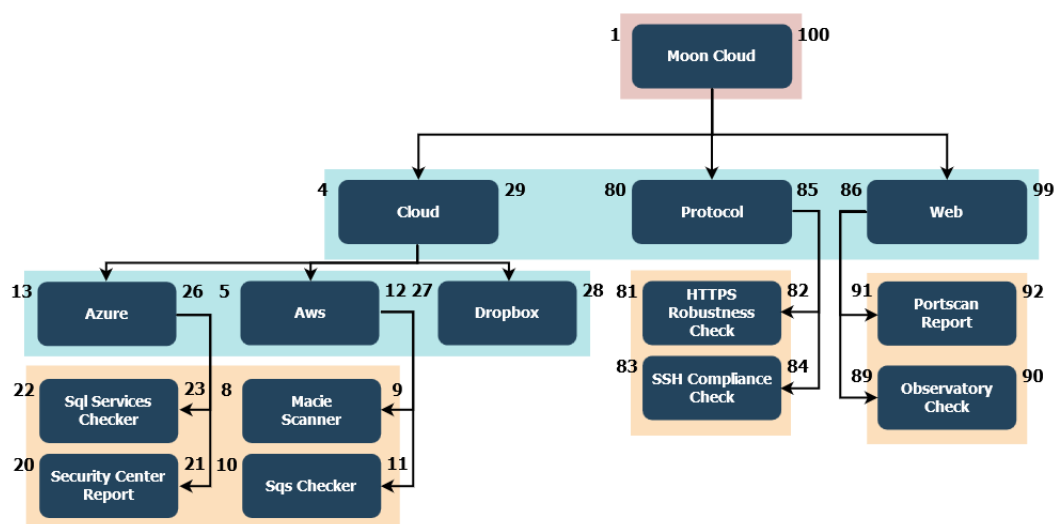


Figura 2.5: Esempio della gestione di dati in modo gerarchico secondo il Nested Set Model, utilizzando dati presi dal database del progetto in questione

si inizia la visita dell'albero partendo da sinistra e continuando verso destra, un livello alla volta, scendendo per ogni nodo i suoi figli, assegnando i valori al campo left, prima di assegnare un valore al campo right, e successivamente si continua verso destra. Questo approccio è chiamato *Modified preorder tree traversal algorithm* (MPTT). A partire da questa tecnica è stata ideata la struttura della tassonomia delle evaluation e dei controlli implementate nella soluzione proposta in questa tesi, con l'ausilio di un package di Python chiamato MPTT.

Più semplicemente se si osserva la parte superiore della Figura 2.4 possiamo notare che la numerazione dei nodi, viene effettuata a partire da container più esterno da sinistra e continua verso destra.

A prima vista questo approccio può sembrare più complicato da comprendere rispetto all'Adjacency List Model, ma quest'ultimo è molto più veloce quando si vuole recuperare i nodi, visto che basta una query, mentre è più lento per operazioni di aggiornamento e cancellazione dei nodi; in quest'ultimo il grado di complicatezza dell'operazione è determinato dal nodo che si vuole cancellare, a partire dal caso più semplice, il nodo foglia (un nodo senza figli) fino al caso più complicato, quando si vuole cancellare il nodo radice.

2.4 Sistemi di raccomandazione

Un sistema di raccomandazione (*Recommendation System*) è un sistema che consiglia ad utenti uno o più item esistenti in un database. L'*item* è un qualsiasi cosa di interesse agli utenti, come prodotti, libri o giornali. Quando si eseguire una raccomandazione si hanno delle aspettative che l'item raccomandato possa essere tra quelli di maggiore interesse; in altre parole, devono essere in accordo con i gusti degli utenti.

Oggigiorno si possono trovare due principali trend di sistemi di raccomandazione:

Content-based filtering (CBF): un item viene raccomandato ad un utente se esso è simile ad altri item di interesse o piaciuti in passato, vengo presi in considerazione prima gli item con alte valutazioni o quelli molto utilizzati. Ogni item ha associate delle informazioni che lo descrivono, questo insieme di dati viene definito metadati, e sono fondamentali per il processo di raccomandazione.

Collaborative filtering (CF): un item viene raccomandato ad un utente se i suoi vicini (altri utenti simili) sono interessati a quello stesso item.

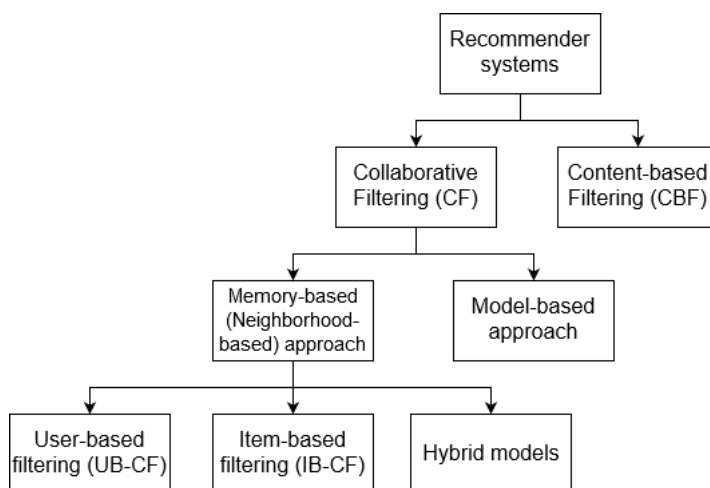


Figura 2.6: Categorizzazione generale dei sistemi di raccomandazione

Entrambi gli approcci (CBF e CF) hanno i loro punti di forza e di debolezza. Il primo algoritmo si focalizza sul contenuto degli item e sugli interessi del singolo utente e propone item differenti a utenti differenti, questo significa che ogni utente può ricevere raccomandazioni uniche; e questo è un

vantaggio. Tuttavia la più grande limitazione del CBF è il fatto di non poter determinare se un utente è interessato ad un item in modo implicito, perchè analizza solamente direttamente i metadati del prodotto e non considera gli interessi di altri utenti, i quali potrebbero suggerire item che non verrebbero notati con questo approccio. Per quanto riguarda il CF, nel caso siano presenti molti contenuti e proprietà associati agli item allora il sistema CF consuma molte risorse e tempo per poter analizzarli, nel contempo a questo algoritmo non interessano queste informazioni. Una raccomandazione viene fatta sulla base delle valutazioni degli utenti per gli item, o sugli usi che gli utenti fanno degli item e questo è il suo punto di forza perchè non si trova a dover analizzare item ricchi di informazioni. Allo stesso tempo è anche il suo punto debole, perchè può portare a suggerimenti che potrebbero essere considerati poco adatti sulla base della poca relazione con i profili di alcuni utenti. Questo problema è accentuato quando sono presenti nel database molti item che non hanno valutazioni o non sono stati mai usati dagli utenti. [4]

Un sistema di raccomandazione filtra i dati usando differenti algoritmi e raccomanda gli item più rilevanti agli utenti, attraverso un procedimento a 3 fasi:

Raccolta di dati : questo è il primo step e anche quello più importante per poter costruire un sistema di raccomandazione che produca risultati rilevanti e consistenti. I dati possono essere raccolti in due modi: esplicitamente, cioè attraverso la raccolta diretta di informazioni fornite dagli utenti, ad esempio le valutazioni di un prodotto; mentre attraverso l'approccio implicito, vengono raccolti dati che non sono prodotti in modo intenzionale dall'utente ma ottenuti dai costanti flussi di dati come la cronologia di ricerca, i click effettuati, lo storico degli ordini, etc.

Memorizzazione di dati : la quantità di dati definisce quanto efficace un modello di raccomandazione possa diventare. Ad esempio, in un sistema di raccomandazione per film, maggiori sono le valutazioni fornite dagli utenti, e migliore sarà il sistema di raccomandazione per gli altri utenti. Il tipo di dati che si vuole raccogliere determina anche il supporto di memorizzazione più adatto.

Filtraggio dei dati : dopo la fase di raccolta e memorizzazione dei dati, essi vanno filtrati per poter estrarre le informazioni rilevanti per poter effettuare le raccomandazioni finali, e sono già disponibili diversi algoritmi che semplificano quest'ultima fase.

I sistemi di raccomandazione possono essere suddivisi nelle seguenti categorie, ma spesso si preferisce degli approcci ibridi, delle combinazioni di sistemi di raccomandazione basati sul Contenuto (*Content-based filtering*) e di quelli Collaborativi (*Collaborative filtering*) in modo da essere più efficaci e sfruttare i pregi di entrambi gli approcci.

2.4.1 Content-based filtering

Un Content-based filtering (acronimo CBF) è un sistema di raccomandazione in cui vengono suggeriti, rispetto ad un item (oggetti o prodotti), quelli più simili, il confronto viene effettuato sfruttando i metadati, come il genere, una descrizione, uno o più autori, la categoria di appartenenza, etc.; l'idea base che si trova dietro questi sistemi, si basa sul fatto che se ad un utente piace o interessa un particolare item allora gli piaceranno anche altri con caratteristiche o proprietà simili.

Questo algoritmo suggerisce prodotti che piacevano all'utente nel passato ed è limitato a item dello stesso tipo. Un Content-based recommender fa riferimento a quegli approcci, che provvedono raccomandazioni comparando la rappresentazione del contenuto che descrive un item e la rappresentazione del contenuto dell'item interessato dall'utente.

Questi metodi sono usati quando si conoscono a priori i metadati sugli item che si vuole suggerire, ma nulla sugli utenti. In questo sistema, delle *keyword* sono utilizzate per caratterizzare gli item e un profilo dell'utente è costruito per memorizzare quali item sono di suo interesse. In altre parole, questi algoritmi cercano di raccomandare quello che l'utente ha valutato positivamente o usato nel passato e sta esaminando nel presente. La costruzione del profilo dell'utente, spesso temporaneo, non viene basata su un modulo di registrazione che l'utente stesso deve compilare, ma su informazioni lasciate indirettamente dall'utente, le quali possono essere: i prodotti che ha maggiormente cercato e acquistato, quelli che sono stati inseriti nella lista dei desideri, etc.. Più precisamente, tra vari item candidati da raccomandare all'utente si passa per un processo di confronto con gli item piaciuti dall'utente e gli item migliori vengono suggeriti.

2.4.2 Collaborative filtering

I Filtri Collaborativi (*Collaborative filtering*, acronimo CF) per poter funzionare come prima cosa costruiscono un database di preferenze degli utenti sulla base di un insieme di item (o prodotti), che a loro volta possono essere presenti un database, sfruttano tecniche di analisi dei dati per risolvere il problema di aiutare gli utenti a trovare gli item che gli potrebbero piacere

producendo, eventualmente una lista dei top-N item da raccomandare per un dato utente. Un utente è sottoposto ad un processo di matching all'interno del database per scoprire quali sono i possibili *neighbors*, che corrispondono ai possibili utenti aventi storicamente delle preferenze in comune al lui. A questo punto gli item maggiormente preferiti dai neighbors sono raccomandati all'utente visto che potrebbero essere di suo interesse.

Questi sistemi tentano di predire la valutazione o la preferenza che un utente darebbe a un item basandosi su preferenze date da altri utenti, queste ultime possono essere ottenute o in modo esplicito dagli utenti o tramite misurazioni implicite. Inoltre i Filtri Collaborativi non richiedono l'uso di metadati associati agli item come nella loro controparte, i filtri Content-based.

Tuttavia, restano ancora oggi alcune sfide significative a cui sono sottoposti i sistemi di raccomandazione basati su Filtri Collaborativi.

Il primo obiettivo è quello di migliorare la scalabilità degli algoritmi di Filtri Collaborativi; questi algoritmi sono in grado di cercare anche diecimila di potenziali neighbors (utenti simili) in tempo reale, ma la richiesta dei sistemi moderni è di cercare dieci milioni di potenziali neighbors, per questo motivo possono nascere problemi di performance con i singoli utenti quando essi hanno molte informazioni.

Il secondo obiettivo è quello di migliorare la qualità dei sistemi di raccomandazione per gli utenti. Gli utenti vogliono raccomandazioni di cui possono fidarsi e che possono aiutarli a trovare item che potrebbero essere di loro gusto.

Per certi versi questi due obiettivi sono in conflitto tra di loro e per ottenere dei risultati validi e di una certa importanza è necessario trattarli in contemporanea perchè aumentare solamente la scalabilità diminuirebbe la sua qualità e viceversa. [8]

Il principale modello di Filtri Collaborativi studiato in questo elaborato e approfondito nel capitolo successivo, è il metodo definito come *Memory-based* e il vantaggio di utilizzare queste tecniche sta nel fatto di essere semplici da implementare e i risultati ottenuti sono altrettanto semplici da interpretare; mentre ci possono essere anche Filtri Collaborativi che sfruttano metodi *Model-based* che si basano sulla fattorizzazione di matrici e sono molto più funzionali per gestire il problema della parsità dei dati. Questi ultimi sono sviluppati usando algoritmi di data mining e machine learning per predire le valutazioni di utenti su item senza valutazioni, tentando di comprimere grandi database in un modello ed effettuare il processo di raccomandazione

applicando dei meccanismi di riferimento all'interno di questo modello. I CF Model-based possono rispondere alle richieste degli utenti istantaneamente. [4]

2.4.3 Cold Start problem

Cosa succederebbe se un nuovo utente o un nuovo item venisse aggiunto al database? Questa situazione è chiamata *Cold Start*. Ci possono essere due tipi di Cold Start:

Visitor Cold Start : si verifica quando un nuovo utente è stato aggiunto al database, visto che non c'è alcuno storico relativo ad esso, il sistema non sa le sue preferenze; per questo motivo diventa molto più difficile raccomandare prodotti a quel particolare utente. Per risolvere questo problema, a livello teorico, si potrebbe applicare un procedimento di raccomandazione basata sulla popolarità dei prodotti, ma solo una volta che si è venuti a conoscenza delle preferenze dell'utente, sarà possibile generare delle raccomandazioni più precise e adeguate alle sue esigenze.

Item Cold Start : si verifica quando un nuovo item viene inserito nel sistema. L'azione dell'utente è quella più importante per determinare il valore di questo item; maggiore l'interazione un item riceve, più è facile che venga raccomandato all'utente giusto.

Capitolo 3

Collaborative filtering

In questo capitolo verranno approfonditi gli algoritmi di raccomandazione implementati nella soluzione, mostrando le porzioni di codice e spiegando i vari passaggi che portano ad ottenere delle raccomandazioni.

3.1 Memory-based

I Filtri Collaborativi Memory-based sono stati introdotti per via delle osservazioni che vennero fatte dalla comunità, dicenddo che gli utenti si fidano maggiormente delle raccomandazioni di altri che la pensano allo stesso modo. Questi metodi mirano a calcolare le relazioni tra utenti e item attraverso lo schema dei vicini che identifica sia coppie di item che tendono ad essere usati insieme o hanno un grado di similarità alto o utenti con uno storico di item usati simile. [6] Questi approcci divennero molto famosi grazie alla loro semplicità di implementazione, molto intuitivi, non necessitano di operazioni di training sui dati e regolazione di molti parametri, inoltre l'utente può capire la ragione che sta dietro ogni raccomandazione.

Questa categoria di sistemi di raccomandazione sono definiti anche *Neighborhood-based* e possono essere ulteriormente classificati in due sottocategorie:

3.1.1 User-based filtering

Questo sistema, definiti anche con l'acronimo UB-CF (*User-based Collaborative Filter*) basa tutto il suo funzionamento sulla comunità di utenti, maggiore è la sua dimensione e l'attività degli utenti con item o servizi e migliori potranno essere le raccomandazioni. Questo algoritmo fornisce dei suggerimenti ad un utente sulla base di uno o più vicini, e la similarità tra utenti può essere determinata sulla base degli item che l'utente ha utilizzato o valutato.

Molti di questi approcci possono essere generalizzati dall'algoritmo definito dai seguenti step:

1. Specificare qual'è l'utente a cui si vuole applicare l'algoritmo di raccomandazione e recuperare quali utenti possono avere dato valutazioni o usato item simili al utente target. Piuttosto che recuperare tutti gli utenti, per velocizzare l'esecuzione dell'algoritmo, è possibile selezionare soltanto un gruppo di utenti in modo casuale oppure associare dei valori di similarità tra tutti gli utenti e confrontando questi valori con quello dell'utente target, selezionare i relativi utenti che superano una soglia scelta, oppure utilizzare tecniche di clustering.
2. Estrarre quegli item a cui l'utente target non ha mai interagito e per questo motivo gli possono interessare, e mostrarli sottoforma di raccomandazioni.

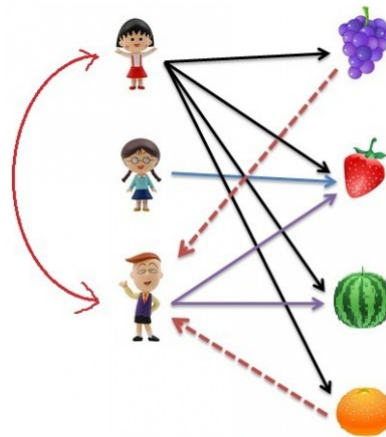


Figura 3.1: Esempio di applicazione di un sistema di raccomandazione User-based

Questi approcci sono facilmente implementabili, indipendenti dal contesto in cui sono applicati e possono essere più accurati rispetto a tecniche basate sul Content-based filtering; dall'altra parte all'aumentare del numero di utenti che vado a considerare per fare le raccomandazioni migliore è la precisione di questo processo ma anche è maggiore il costo in termini di tempo.

3.1.2 Item-based filtering

Quando viene applicato per milioni di utenti e item, l'algoritmo UB-CF non è molto efficiente, per via della complessa computazione della ricerca di utenti

simili; così in alternativa è stato introdotto l'algoritmo di filtraggio Item-based, definito anche IB-CF (*Item-based Collaborative Filter*) dove piuttosto che effettuare il confronto tra utenti simili, viene fatto un confronto tra gli item dell'utente a cui si vuole raccomandare e i possibili item simili.

Questi sistemi sono estremamente simili ai sistemi di raccomandazione Content-based, e identificano item simili in base a come utenti gli hanno usati nel passato. [8]

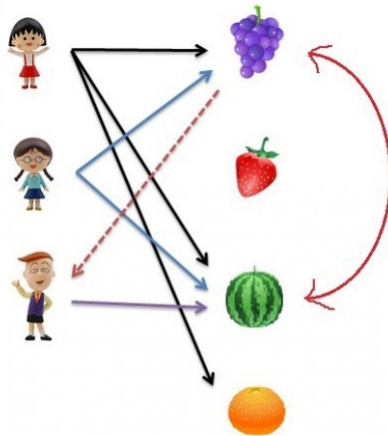


Figura 3.2: Esempio di applicazione di un sistema di raccomandazione Item-based

3.1.3 Hybrid filtering

Capitolo 4

Descrizione della soluzione

In questo capitolo verrà approfondito l'aspetto puramente pratico e le fasi che hanno portato alla realizzazione della soluzione; inoltre verranno mostrate le applicazioni pratiche degli aspetti teorici enunciati nei capitoli precedenti, unitamente alle difficoltà principali incontrate.

Prima di poter costruire il sistema di raccomandazione proposto in questa tesi, sono state eseguite delle operazioni preliminari per poter impostare il progetto di Django e la relativa applicazione che implementerà effettivamente la soluzione. Dalla Figura 4.1 è possibile osservare come viene strutturato il progetto, le relative cartelle e file.

Il primo passo compiuto e fondamentale è stato quello di disegnare una base di dati da cui partire per la realizzazione degli algoritmi proposti. In generale Moon Cloud possiede una struttura delle Evaluation ad albero, quindi anche di conseguenza anche le tabelle del database rispecchiano questa struttura, sulla base delle considerazioni sulle tecniche adottate sono state fatte nei capitoli precedenti.

Per implementare un modified pre-order trasversal tree in Django, si è fatto uso del package MPTT, come detto in precedenza, questa tecnica è usata per memorizzare dati gerarchici in un database, puntando all'efficienza nelle operazioni di recupero dei dati e scendendo a compromessi per quanto riguarda le operazioni di inserimento e spostamento dei nodi all'interno della struttura. Grazie all'utilità di questa utility ai Models

Si è pensato di introdurre una tabella Evaluation, che contenga l'insieme di Evaluation che un utente può eseguire per un certo Target, una tabella Control, che rappresenta l'insieme di controlli che possono comporre un Evaluation, una tabella TargetType, che definisce i tipi di Target supportati da MoonCloud su cui è possibile effettuare i monitoraggi, la tabella Target, raccoglie quali Target un utente ha inserito e possono essere più di uno.

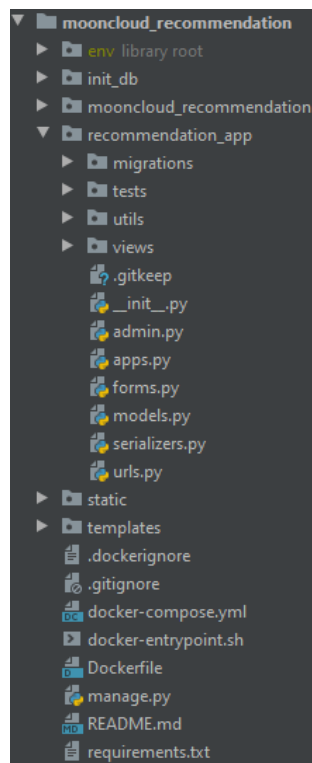


Figura 4.1: Struttura dell'applicativo sfruttando il framework proposto da Django

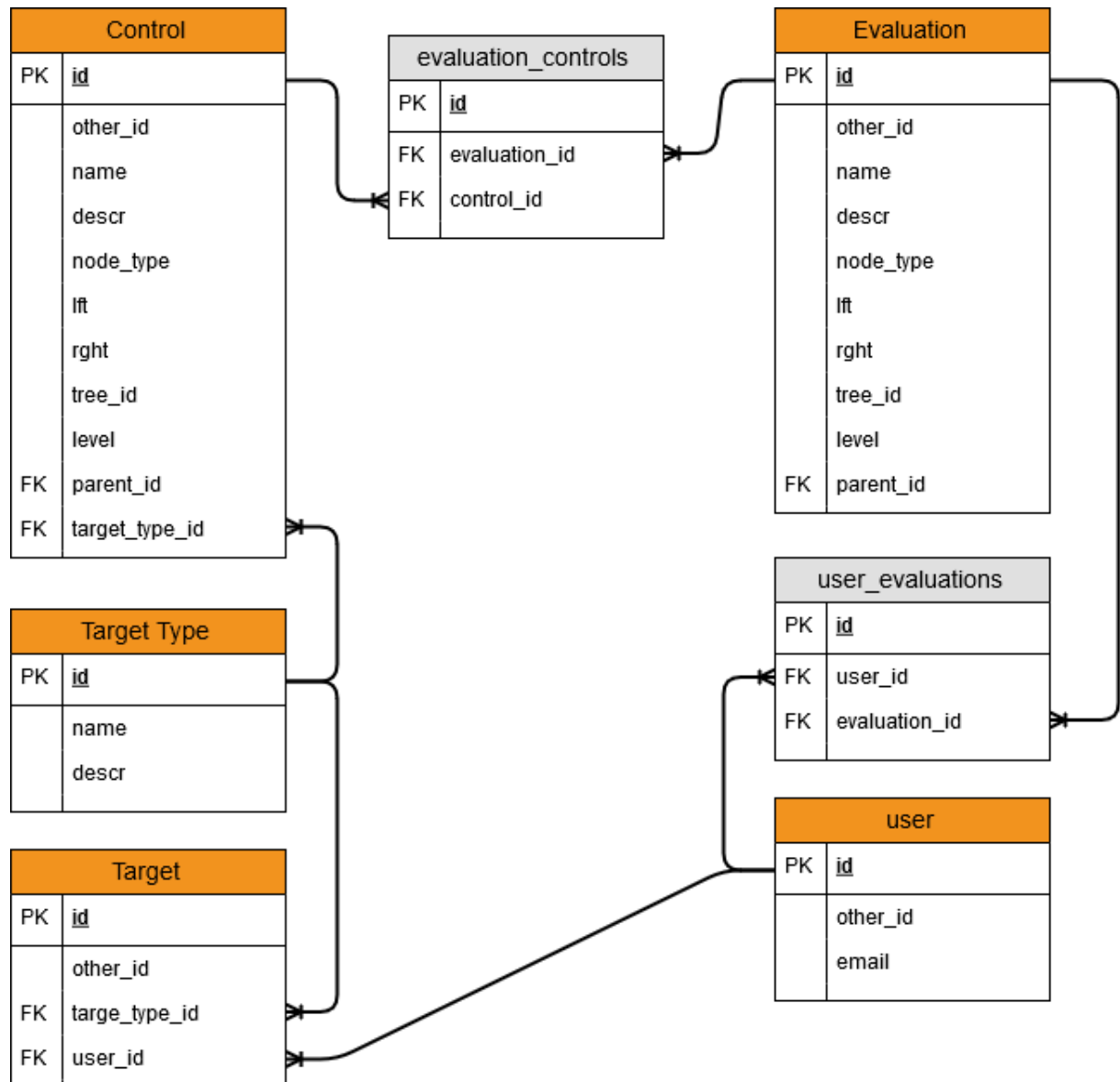


Figura 4.2: Struttura del database

Come annunciato nei capitoli precedenti per procedere alla costruzione di un sistema di raccomandazione bisogna avere a disposizione una base di dati solida da cui attingere tutte le informazioni; ed è proprio questo il primo passo che è stato seguito,

Capitolo 5

Conclusioni

Presentazione dei risultati e conclusioni: la presentazione dei risultati dovrebbe consistere in una descrizione tecnica dei risultati raggiunti, unitamente ad un commento critico e ad un'analisi della rispondenza agli obiettivi iniziali (si consiglia pertanto di motivare la rilevanza dei risultati e l'eventuale scostamento dagli obiettivi iniziali). La sezione relativa ai risultati dovrebbe infine contenere una sintesi critica e un giudizio sull'esperienza effettuata, che renda conto di aspetti positivi e negativi per il tirocinante e per l'ente ospitante, del valore formativo, professionale e umano, così via.

Bibliografia

- [1] M. Anisetti et al. «A semi-automatic and trustworthy scheme for continuous cloud service certification». In: *IEEE TRANSACTIONS ON SERVICES COMPUTING* (2017). DOI: 10.1109/TSC.2017.2657505.
- [2] M. Anisetti et al. «Moon Cloud: A Cloud Platform for ICT Security Governance». In: (dic. 2018), pp. 1–7. DOI: 10.1109/GLOCOM.2018.8647247.
- [3] *Django documentation*. <https://docs.djangoproject.com/en/2.2/>. 2019.
- [4] Minh-Phung Do, Dung Nguyen e Academic Network of Loc Nguyen. «Model-based approach for Collaborative Filtering». In: ago. 2010.
- [5] *MDN Django documentation*. <https://developer.mozilla.org/it/docs/Learn/Server-side/Django/>. 2019.
- [6] Miquel Montaner, Beatriz López e Josep Lluís de la Rosa. «A Taxonomy of Recommender Agents on the Internet». In: *Artificial Intelligence Review* 19.4 (giu. 2003), pp. 285–330. ISSN: 1573-7462. DOI: 10.1023/A:1022850703159. URL: <https://doi.org/10.1023/A:1022850703159>.
- [7] *Python 3.7 documentation*. <https://docs.python.org/3.7/>. 2019.
- [8] Badrul Sarwar et al. «Item-based Collaborative Filtering Recommendation Algorithms». In: WWW '01 (2001), pp. 285–295. DOI: 10.1145/371920.372071. URL: <http://doi.acm.org/10.1145/371920.372071>.