Laboratorio Basi di Dati

Anno accademico 2023/2024

Bortuzzo Francesco
francesco.bortuzzo@spes.uniud.it 157430

Indice

1. Introduzione e obbiettivo del progetto

1.1. Introduzione al batterio Legionella

Il batterio Legionella è un bacillo gram-negativo aerobio, non mobile, che si sviluppa in ambienti acquatici e umidi sia naturali, come acque sorgive, termali, di fiumi o di laghi, che artificiali, come tubature, serbatoi, fontane e piscine. La Legionella è in grado di sopravvivere in una vasta gamma di condizioni ambientali, tra cui temperature comprese tra 20 e 45 gradi Celsius, pH neutro o leggermente alcalino e presenza di nutrienti organici. Il genere comprende 62 specie diverse, suddivise in 71 sierotipi, di cui circa 20 sono patogeni per l'uomo. La specie più comune è la Legionella pneumophila, responsabile della maggior parte dei casi di legionellosi.

La legionellosi è una malattia infettiva che si manifesta con sintomi simili a quelli dell'influenza, come febbre, tosse, dolori muscolari e mal di testa. La malattia può essere contratta inalando aerosol contenenti il batterio, come ad esempio le goccioline d'acqua presenti negli impianti di condizionamento e nei sistemi di riscaldamento. È dunque di fondamentale importanza monitorare la diffusione di questo batterio negli ambienti umidi e acquatici, al fine di prevenire la malattia e proteggere la salute pubblica. Particolare attenzione deve essere rivolta alle strutture ospedaliere e alle strutture termali e alberghiere, che per loro natura rappresentano ambienti a rischio di diffusione del batterio.

La malattia può manifestarsi in due forme: la forma più comune è la legionellosi non polmonare, che si manifesta con sintomi influenzali e può essere facilmente curata con antibiotici; la forma più grave è la legionellosi polmonare, che si manifesta con sintomi simili a quelli della polmonite e può portare a complicazioni gravi, come la polmonite atipica o il decesso. Il primo episodio registrato, da cui deriva il nome del batterio, risale al 1976 quando un'epidemia colpì i partecipanti di un raduno della Legione Americana in un famoso hotel di Philadelphia. In quell'occasione, 224 persone contrassero una forma di polmonite al tempo sconosciuta, risultata fatale per 32 di esse. Le indagini effettuate in tale occasione identificarono nel sistema di aria condizionata dell'albergo il mezzo di propagazione del batterio.

1.2. Legionella in Friuli Venezia Giulia

Nell'Unione Europea, la raccolta di dati relativi alla presenza del batterio è effettuata dall'ECDC[1], un ente istituito nel 2005. Nel nostro Paese, invece, questa attività è svolta da diversi enti e istituzioni. Un contributo significativo proviene dal Ministero della Salute e dai vari organismi che costituiscono il SNPA[2], di cui fa parte l'ARPA FVG[3]. I dati raccolti sono utilizzati per valutare il rischio di diffusione del batterio e adottare misure di prevenzione e controllo.

Proprio l'ARPA FVG ha condotto diverse indagini sulla presenza di Legionella nei sistemi di aduzione e conservazione dell'acqua della regione e ha pubblicato i risultati in vari report. Ad esempio nel 2019 ARPA FVG e Università degli studi di Udine hanno collaborato per la pubblicazione di un articolo(Felice et al., 2019) riguardante la presenza di legionella nei sistemi di raccolta e distribuzione dell' acqua nella regione, in un periodo

[1]Centro europeo per la prevenzione e il controllo delle malattie

[2]Sistema Nazionale per la Protezione dell'Ambiente

[3]Agenzia Regionale per la Protezione dell'Ambiente Friuli Venezia Giulia

di sedici anni, dal 2002 al 2017, durante il quale sono stati raccolti e analizzati 20.319 campioni in 3.983 indagini ambientali.
I risultati derivati dallo studio, riferiti alle indagini non clinciche[4], hanno evidenziato che la presenza di Legionella è diffusa sopratutto nei cluster di impianti termali (il batterio è stato individuato nel 57,8% dei siti indagati) e ospedalieri (nel 50,8% delle strutture è stata riscontrata la legionella almeno una volta) con picchi dei campioni positivi soprattutto nei mesi che segnano l'inizio del periodo autunnale. Sebbene la presenza del batterio abbia mostrato un trend crescente nel periodo considerato, si è osservato un forte incremento tra la seconda metà del 2006 e l'inizio del 2009, seguito da un calo fino al 2013 e un nuovo aumento negli anni successivi. Questo andamento evidenzia la necessità di monitorare costantemente la presenza di Legionella e di adottare misure di prevenzione e controllo per evitare la diffusione del batterio e ,conseguentemente, ridurre il rischio di nuove epidemie.

1.3. Obiettivo del progetto

Al fine di analizzare i dati acquisiti e studiare la diffusione del batterio, è opportuno utilizzare un sistema informativo che permetta di memorizzare, gestire e interrogare i dati in modo efficiente. Tuttavia, in regione, il vasto numero di dati già raccolti non è stato organizzato in modo efficace e pertanto non è possibile effettuare ricerche senza prima ristrutturare e connettere i vari dataset.
In questo contesto, i sistemi di basi di dati giocano un ruolo fondamentale, in quanto permettono di memorizzare grandi quantità di dati e di effettuare ricerche complesse in modo rapido ed efficiente. In particolare, i sistemi di basi di dati a grafo sembrano particolarmente adatti per la modellazione e l'analisi di dati complessi, come quelli relativi alla diffusione della legionella poichè permettono di rappresentare le relazioni tra i dati in modo naturale e di effettuare ricerche complesse in modo efficiente.

Questo documento mira a condurre un'analisi critica di un database relazionale nell'ambito delineato e a proporre un'alternativa attraverso l'impiego di un database a grafo. In particolare, si illustrerà il processo di modellazione, creazione e popolamento di un database a grafo, utilizzando Neo4j, per l'analisi dei dati sulla diffusione della legionella nella nostra regione.

[4]eseguite routinariamente nell'ambito del piano regionale di sorveglianza ambientale

2. Analisi critica del database relazionale

Questa sezione è dedicata all'analisi critica di un database relazionale utilizzato per memorizzare i dati relativi alla diffusione della legionella. Il database oggetto di analisi è stato progettato dal dottor Dario Garlatti nell'ambito della sua tesi di laurea triennale in informatica, intitolata "Base di dati e applicazione web per il monitoraggio del batterio della legionella".

2.1. Analisi dei requisiti

Prima di procedere con lo studio del database, è necessario definire i requisiti del sistema informativo. Questi requisiti sono di natura qualitativa e descrivono le caratteristiche che il sistema deve possedere per soddisfare le esigenze degli utenti e degli stakeholder. Nel nostro contesto, i requisiti riguardano l'intera fase di acquisizione dei dati relativi alle e indagini ambientali per il monitoraggio del batterio legionella.

2.1.1. Requisiti non strutturati

In particolare, i requisiti non strutturati del sistema informativo sono i seguenti:

Il sistema deve permettere la registrazione di indagini ambientali relative alla presenza di Legionella nei sistemi di aduzione e conservazione dell'acqua.
Un'indagine ambientale è caratterizzata dal tipo, dalla data e dal sito presso cui viene condotta ed è associata al richiedente, qualora si tratti di un'indagine di follow-up.
Un sito è caratterizzato da un indirizzo e da una categoria.
L'indagine consiste nel prelievo di campioni per analizzarli alla ricerca del batterio Legionella. Ognuno dei campioni prelevati è associato a una specifica indagine, è caratterizzato dal punto di prelievo all'interno del sito ed è identificato da un codice univoco. Tutti i campioni prelevati devono essere sottoposti a diverse analisi per determinare la presenza o l'assenza di Legionella:

1. PCR[5] qualitativa: permette di identificare la presenza del DNA di Legionella nei campioni prelevati.
2. PCR quantitativa: permette di quantificare la quantità di Legionella presente nei campioni prelevati (in µg/l.).
3. Analise colturale: permette di isolare e identificare le unità formanti colonia UFC_L e, nel caso in cui il campione risulti positivo al batterio, di determinare il sierogruppo.

2.1.2. Requisiti strutturati

I requisiti strutturati del sistema informativo sono i seguenti:

Frasi riguardanti l'indagine ambientale

L'indagine ambientale è caratterizzata dal tipo, dalla data e dal sito presso cui viene condotta ed è associata al richiedente, qualora si tratti di un'indagine di follow-up. L'indagine consiste nel prelievo di campioni per analizzarli alla ricerca della presenza di Legionella.

Frasi riguardanti i campioni

[5]Polymerase Chain Reaction, tecnica che consiste nell'amplificazione dei frammenti di acidi nucleici

Ognuno dei campioni prelevati deve essere associato a una specifica indagine ed è identificato da un codice univoco. Tutti i campioni prelevati devono essere sottoposti a diverse analisi per determinare la presenza o l'assenza di Legionella.

Frasi riguardanti le analisi

Tutti i campioni prelevati devono essere sottoposti a diverse analisi per determinare la presenza o l'assenza di Legionella:

1. PCR qualitativa: permette di identificare la presenza del DNA di Legionella nei campioni prelevati.
2. PCR quantitativa: permette di quantificare la quantità di Legionella presente nei campioni prelevati (in µg/l.).
3. Analise colturale: permette di isolare e identificare le unità formanti colonia UFC_L e, nel caso in cui il campione risulti positivo al batterio, di determinare il sierogruppo.

Frasi riguardanti i siti

Un sito è caratterizzato da un indirizzo e da una categoria.

2.2. Schema relazionale

Per rappresentare i dati relativi alle indagini ambientali e alle analisi effettuate sui campioni prelevati, è stato progettato il seguente schema relazionale.

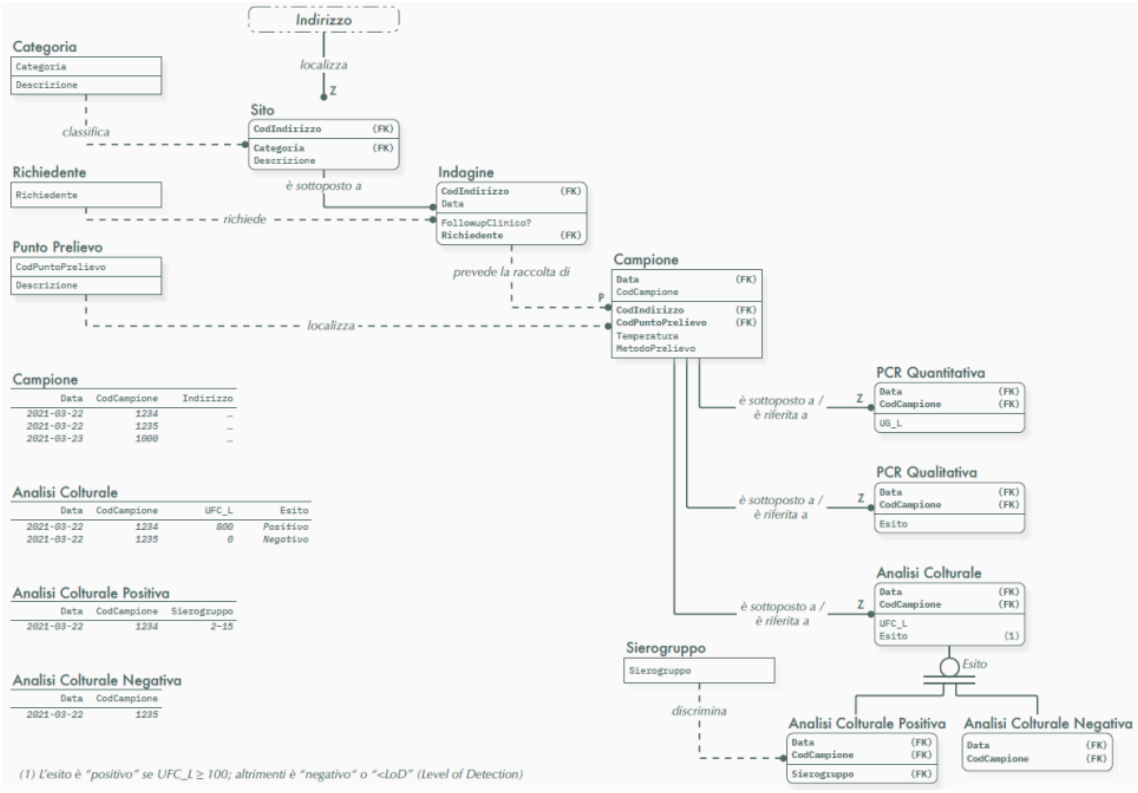

Figura 1: Diagramma ER

2.3. Glossario

Per facilitare la comprensione dello schema relazionale, è stato redatto un glossario contenente le definizioni dei termini tecnici utilizzati nel documento.

Termine	**Descrizione**
Campione	Piccola quantità di acqua prelevata durante un'indagine ambientale
Categoria	Categoria di appartenenza di un sito (es. ospedaliero termale, alberghiero)
Indagine Ambientale	Indagine condotta per verificare la presenza di Legionella in un sito
PCR Qualitativa	Esito dell'analisi che permette di identificare la presenza del DNA di Legionella
PCR Quantitativa	Esito dell'analisi che permette di quantificare la quantità di Legionella presente nei campioni
Analisi Colturale	Esito dell'analisi che permette di isolare e identificare le unità formanti colonia UFC_L
Richiedente	Ente che richiede l'indagine ambientale
Sieogruppo	Gruppo di sierotipi di Legionella. Il laboratorio ARPA distingue tre sierotipi: sierotipo 1, sierotipo 2-15 e sierotipo sp
Sito	Struttura presso cui viene condotta un'indagine ambientale
Indirizzo	indirizzo del sito presso cui viene condotta un'indagine ambientale. Segue modello ANNCSU[6]
Punto di prelievo	Punto all'interno del sito in cui è stato prelevato un campione di acqua

Tabella 1: Glossario delle entità

[6]Archivio Nazionale dei Numeri Civici e delle Strade Urbane

Termine	Descrizione
Esito	Esito qualitativo dell'analisi di un campione di acqua prelevato durante un'indagine ambientale
FollowUp Clinico	Dominio booleano che indica il tipo dell'indagine, ovvero se l'indagine è di follow-up clinico oppure se è avviata nell'ambito del normale piano di monitoraggio.
MetodoPrelievo	Dominio di due valori che spcifica il metodo utilizzato per prelevare i campioni di acqua durante un'indagine ambientale (istantaneo o quantitativo)
UFC_L	valore numerico espresso in UFC per litro
UG_L	valore numerico espresso in µg/l
Temperatura	Dominio di due valori (caldo o freddo) che esprime la temperatura dell'acqua in cui è stato prelevato un campione
Sierotipo	Dominio di tre valori che definisce il gruppo di sierotipi di Legionella (sierotipo 1, sierotipo 2-15, sierotipo sp)

Tabella 2: Glossario dei domini

2.4. Analisi critica del database

L'analisi critica del database relazionale è finalizzata a valutare i punti di forza e di debolezza del sistema informativo progettato dal dottor Dario Garlatti. In particolare, si analizzeranno i seguenti aspetti:

2.4.1. Struttura del database

La struttura del database è stata progettata in modo da rappresentare le entità coinvolte nel processo di monitoraggio della legionella e le relazioni tra di esse. Tuttavia, la struttura del database presenta alcune criticità, tra cui:

1. Ridondanza dei dati: A causa delle relazioni molti a molti nei database relazionali, alcune informazioni sono duplicate in più tabelle, aumentando la complessità del sistema e il rischio di errori. Per garantire la consistenza dei dati, è necessario implementare vincoli di integrità referenziale e procedure di aggiornamento specifiche.
2. Schema poco flessibile: Lo schema del database è poco flessibile e non permette di gestire facilmente nuove entità o relazioni tra le entità.
3. Complessità della gestione dei vincoli di integrità referenziale: Non solo la ridondanza dei dati, ma anche i legami indiretti tra alcune tabelle rendono difficile la gestione dei vincoli di integrità referenziale. Ad esempio, per garantire la consistenza dei dati registrati nelle tabelle PCR Qualitativa e PCR Quantitativa, è necessario implementare un vincolo di integrità che assicuri che a un campione positivo sia associato un valore UG_L positivo.

2.4.2. Interrogazioni

Le relazioni tra le entità coinvolte nel processo di monitoraggio della legionella sono complesse e possono rendere difficile l'interrogazione del database e l'estrazione di informazioni significative. Ad esempio, per estrarre il livello di contaminazione dei campioni positivi è necessario effettuare una serie di join tra le tabelle coinvolte, aumentando esponenzialmente la complessità delle interrogazioni.

2.5. Conclusioni

L'analisi ha evidenziato alcune criticità nella struttura del sistema informativo relazionale. In particolare, la ridondanza dei dati, la rigidezza dello schema, la complessità della gestione dei vincoli di integrità referenziale e la complessità delle interrogazioni possono rappresentare dei limiti per l'efficace gestione e analisi dei dati relativi alla diffusione della legionella.

Al fine di superare queste criticità e migliorare l'efficienza del sistema informativo, si propone di implementare un database a grafo per la memorizzazione e l'analisi dei dati sulla diffusione della legionella. In particolare, si utilizzerà Neo4j, un database a grafo open source, per modellare, creare e popolare il database e per effettuare interrogazioni complesse in modo efficiente.

3. Neo4j

(Robinson et al., 2015)

3.1. Specifica di implementazione dei nodi

3.2. Specifica di implementazione delle relazioni

(cypher)

4. implementazione su base di dati a grafo

4.1. descrizione della struttura dei nodi

4.2. descrizione delle relazioni tra nodi

(introduzione di uno schema generale con nodi e relazioni tra essi)

4.3. implementazione

4.4. operazioni

4.5. popolamento

5. grafici

Bibliografia

Felice, A., Franchi, M., De Martin, S., Vitacolonna, N., Iacumin, L., & Civilini, M. (2019). Environmental surveillance and spatio-temporal analysis of Legionella spp. in a region of northeastern Italy (2002–2017). *PLOS ONE*, *14*(7), e218687. https://doi.org/10.1371/journal.pone.0218687

Robinson, I., Webber, J., & Eifrem, E. (2015). *Graph Databases: New opportunities for connected data*. O'Reilly Media.

6. appunti