



UNIVERSITÀ DEGLI STUDI DI FIRENZE
SCUOLA DI INGEGNERIA - DIPARTIMENTO DI INGEGNERIA
DELL'INFORMAZIONE

Tesi di Laurea Magistrale in Ingegneria Informatica

**PROGETTAZIONE E SVILUPPO DI COMPONENTI
PER LA PIATTAFORMA AIRQINO DEDICATA AL
MONITORAGGIO DELLA QUALITÀ DELL'ARIA**

Candidato

Edoardo D'Angelis

Relatori

Prof. Andrew D. Bagdanov

Prof. Pietro Pala

Correlatori

Walter Nunziati

Alice Cavaliere

Anno Accademico 2021/2022

Abstract

...

Indice

Abstract	i
1 Introduzione	1
1.1 Contesto	1
1.1.1 Descrizione del problema	1
1.1.2 Progetti simili	2
1.2 La piattaforma AirQino	2
1.2.1 Hardware dei sensori	2
1.2.2 Architettura e tecnologie	2
1.2.3 Progetti correlati	2
2 Sviluppi tecnologici	3
2.1 Replica del database di produzione	3
2.1.1 Motivazioni	3
2.1.2 Streaming Replication	3
2.2 Ottimizzazione di query temporali	4
2.2.1 Motivazioni	4
2.2.2 Continuous Aggregates	4
2.2.3 Risultati ottenuti	5

3	Calibrazione	6
3.1	I dati a disposizione	6
3.1.1	Dataset NO2	6
3.1.2	Dataset PM2.5 e PM10	6
3.1.3	Preprocessamento	6
3.2	Regressione	7
3.2.1	Regressione lineare	7
3.2.2	Regressione lineare robusta (Huber)	8
3.2.3	Regressione lineare avanzata (con rimozione di outlier)	8
3.2.4	Regressione Lasso	8
3.2.5	Regressione Ridge	9
3.2.6	Regressione polinomiale	9
3.3	Esperimenti e risultati ottenuti	9
3.3.1	NO2	9
3.3.2	PM2.5	9
3.3.3	PM10	9
3.4	Validazione	10
3.4.1	PM2.5	10
3.4.2	PM10	10
3.5	Discussione	10
4	Interfaccia di calibrazione	11
4.1	Motivazioni	11
4.2	Tecnologie	11
4.2.1	Backend	11
4.2.2	Frontend	11
4.3	Funzionamento	11
4.4	Autenticazione	12

4.5 CI e deploy automatico	12
Conclusioni e sviluppi futuri	14
Bibliografia	15

Capitolo 1

Introduzione

...

1.1 Contesto

Il monitoraggio della qualità dell'aria è una delle attività più importanti per la tutela della salute pubblica. La qualità dell'aria può essere influenzata da molte sorgenti di emissione, tra cui le automobili, le centrali elettriche, gli impianti di riscaldamento e le fabbriche. I principali inquinanti atmosferici sono il biossido di zolfo, gli idrocarburi policiclici aromatici, il monossido di carbonio e gli ozono. Gli effetti dell'inquinamento atmosferico sulla salute sono molteplici e possono essere a breve o a lungo termine. I principali rischi sono l'asma, le malattie cardiovascolari, il cancro e le malattie respiratorie. Il monitoraggio della qualità dell'aria permette di individuare le sorgenti di emissione e di intervenire per ridurre l'inquinamento atmosferico.

1.1.1 Descrizione del problema

...

1.1.2 Progetti simili

- **Airly** (<https://airly.org/>) è una piattaforma che consente di condividere informazioni ambientali in tempo reale, grazie alla quale è possibile monitorare la qualità dell'aria e i livelli di inquinamento;

1.2 La piattaforma AirQino

...

1.2.1 Hardware dei sensori

...

1.2.2 Architettura e tecnologie

...

1.2.3 Progetti correlati

...

Capitolo 2

Sviluppi tecnologici

2.1 Replica del database di produzione

...

2.1.1 Motivazioni

...

2.1.2 Streaming Replication

La streaming replication di PostgreSQL è una funzionalità che consente di replicare i dati in tempo reale da una istanza di PostgreSQL a un'altra. Questo significa che, se si modificano i dati in una delle istanze, questi saranno immediatamente replicati anche nell'altra istanza. La streaming replication di PostgreSQL offre diversi vantaggi:

- maggiore disponibilità dei dati: se una delle istanze di PostgreSQL viene a mancare, i dati saranno comunque disponibili nell'altra istanza;

- maggiore velocità di replica: i dati vengono replicati in tempo reale, senza dover attendere il completamento delle operazioni di replica;
- riduzione del carico sulle risorse: la replica in tempo reale riduce il carico sulle risorse della infrastruttura di storage.

La replica si basa sulle transazioni WAL (Write Ahead Log) e utilizza il protocollo TCP per garantire una connessione sicura tra i server.

2.2 Ottimizzazione di query temporali

...

2.2.1 Motivazioni

...

2.2.2 Continuous Aggregates

I continuous aggregate sono una funzionalità integrata in TimescaleDB che consente di aggregare i dati in tempo reale, senza la necessità di eseguire query aggiuntive. Questa funzionalità utilizza i contatori per tenere traccia dei dati aggregati in tempo reale e fornisce una rappresentazione dei dati aggregati in tempo reale.

I continuous aggregate offrono numerosi vantaggi, tra cui:

- Flessibilità: è possibile aggregare dati in tempo reale in base a qualsiasi criterio desiderato.
- Risparmio di tempo: non è necessario eseguire query aggiuntive per ottenere informazioni aggregate in tempo reale.
- Risparmio di spazio: i dati aggregati in tempo reale occupano meno spazio rispetto ai dati non aggregati.

- Maggiore efficienza: i continuous aggregate sono più efficienti dei query batch per l'aggregazione dei dati in tempo reale.

2.2.3 Risultati ottenuti

...

Capitolo 3

Calibrazione

3.1 I dati a disposizione

...

3.1.1 Dataset NO2

...

3.1.2 Dataset PM2.5 e PM10

...

3.1.3 Preprocessamento

...

3.2 Regressione

La regressione è una tecnica statistica che serve a stimare la relazione esistente tra due o più variabili. In particolare, la regressione permette di individuare il coefficiente di correlazione tra due variabili e di determinare se questa relazione è casuale o no.

Tra le applicazioni principali della regressione ci sono:

- Stima della relazione tra due variabili
- Analisi della relazione tra variabili
- Valutazione dell'influenza di una variabile sulle altre
- Predicting

3.2.1 Regressione lineare

Per stimare la relazione tra due variabili, la regressione lineare utilizza una formula matematica che calcola la media dei valori della prima variabile (Y) in funzione dei valori della seconda variabile (X). La formula della regressione lineare è:

$$Y = a + bX$$

In questa formula, a è la costante di regressione e b è la coefficiente di regressione. La costante di regressione a indica la media dei valori di Y in funzione dei valori di X . Il coefficiente di regressione b indica la relazione tra le due variabili: più è vicino a 1, più le due variabili sono correlate in modo lineare.

Utilizzando la regressione lineare, è possibile stimare la relazione tra due variabili anche in presenza di deviazioni dalla linea.

3.2.2 Regressione lineare robusta (Huber)

La regressione Huber (in inglese Huber regression, anche detta regressione robusta) è una metodologia statistica per la stima dei parametri di un modello lineare, in presenza di outliers.

Il metodo Huber si basa sul principio della massima verosimiglianza, e si propone di ridurre la sensibilità dei parametri alla presenza di outliers. In particolare, la regressione Huber utilizza una funzione di peso, detta funzione di Huber, che tiene conto della variabilità dei dati intorno ai valori centrali.

3.2.3 Regressione lineare avanzata (con rimozione di outlier)

La regressione con rimozione outlier tramite distanza di Cook è una tecnica statistica per ridurre l'influenza degli outliers nei dati di una regressione lineare.

Si basa sul concetto di distanza di Cook, che misura la distanza tra un dato e il valore medio dei dati della stessa variabile. In presenza di outliers, la distanza di Cook aumenta, e quindi questi dati hanno un maggiore impatto sulle stime dei parametri della regressione.

La regressione con rimozione outlier tramite distanza di Cook si basa sull'utilizzo di una funzione di peso, detta funzione di Cook, che tiene conto della distanza di Cook dei dati. La funzione di peso viene utilizzata per ridurre l'influenza degli outliers nei dati della regressione.

3.2.4 Regressione Lasso

La regressione Lasso (in inglese Lasso regression) è una metodologia statistica per la stima dei parametri di un modello lineare, in presenza di

outliers.

Il metodo Lasso si basa sul principio della massima verosimiglianza, e si propone di ridurre la sensibilità dei parametri alla presenza di outliers. In particolare, la regressione Lasso utilizza una funzione di peso, detta funzione di Lasso, che tiene conto della variabilità dei dati intorno ai valori centrali.

3.2.5 Regressione Ridge

...

3.2.6 Regressione polinomiale

La regressione polinomiale è una generalizzazione della regressione lineare, in cui il rapporto tra Y e X non è più una linea retta.

3.3 Esperimenti e risultati ottenuti

...

3.3.1 NO₂

...

3.3.2 PM_{2.5}

...

3.3.3 PM₁₀

...

3.4 Validazione

...

3.4.1 PM2.5

...

3.4.2 PM10

...

3.5 Discussione

...

Capitolo 4

Interfaccia di calibrazione

4.1 Motivazioni

...

4.2 Tecnologie

...

4.2.1 Backend

...

4.2.2 Frontend

...

4.3 Funzionamento

...

4.4 Autenticazione

Keycloak è un'identità federata open source, sviluppata da Red Hat. Può essere utilizzata per gestire l'autenticazione di utenti e servizi in ambienti cloud e on-premise. I principali vantaggi di Keycloak sono la scalabilità, l'affidabilità e la flessibilità.

Keycloak include un server e un agente. L'agente è installato sulle applicazioni che richiedono l'autenticazione, mentre il server gestisce tutte le richieste di autenticazione. Quando un utente tenta di accedere a una applicazione protetta da Keycloak, l'agente verifica se l'utente è autenticato e, in caso affermativo, fornisce le credenziali appropriate all'applicazione.

4.5 CI e deploy automatico

Continuous integration è una metodologia di sviluppo software che prevede il continuo e costante integrazione dei cambiamenti effettuati dai developer all'interno di un codice sorgente.

La continuous integration ha lo scopo di evitare problemi di sincronizzazione tra gli sviluppatori, riducendo il numero di bug rilevati in fase di testing e aumentando la qualità del codice prodotto.

I principali vantaggi della continuous integration sono:

- riduzione del numero di bug rilevati in fase di testing;
- aumento della qualità del codice prodotto;
- maggiore sincronizzazione tra gli sviluppatori;
- minor rischio di collisioni tra i cambiamenti effettuati dagli sviluppatori.

Jenkins è uno strumento open source di continuous integration. Jenkins permette di automatizzare il processo di integrazione dei cambiamenti effet-

tuati dai developer all'interno di un codice sorgente, eseguendo una serie di controlli per verificarne la correttezza.

Jenkins può essere utilizzato per gestire una varietà di progetti, tra cui sviluppo software, testing, build, deployment e automazione dei processi.

Conclusioni e sviluppi futuri

...

Bibliografia