# Sistemi Informativi Evoluti e Big Data – A.A. 2025-2026

# Homework #2 – InfluxDB

**Introduzione**

L'esercizio richiede di analizzare un dataset di crimini registrati dal Los Angeles Police Department (LAPD) tra il 2020 e il 2025, disponibile a questo link:
https://catalog.data.gov/dataset/crime-data-from-2020-to-present
Il dataset è disponibile e scaricabile in formato CVS. Il compito consiste nel caricare i dati in un database InfluxDB, arricchirli con tag e filtri specifici, impostare regole di warning/allarme in base alla gravità e alla frequenza dei crimini, eseguire delle analisi specifiche con grafici e infine valutare le prestazioni delle query effettuate.

**Descrizione del compito**

***1) Caricamento dei dati.*** Caricare i dati relativi ai crimini a partire dal dataset. Ogni evento (riga) del dataset deve essere interpretato come un record di segnalazione, con:

- il numero della segnalazione (DR_NO);
- la data della segnalazione (Date Rptd) e la data ipotizzata in cui si è verificato il crimine (DATE OCC); i timestamp sono sempre allineati alle ore 12:00:00, pertanto la granularità temporale da adottare è quella giornaliera;
- il distretto (Rpt Dist No) e l'area (con ID e nome) in cui si è verificato il crimine;
- la classificazione del crimine (Part 1-2); la Parte I include i reati più gravi, come omicidio, rapina, furto con scasso e aggressione aggravata; la Parte II comprende invece i reati di minore gravità, come frode, vandalismo e aggressione semplice;
- il tipo di crimine (identificato da un codice, Crm Cd, e da una descrizione, Crm Cd Desc;
- attributi della vittima, se disponibili (età, genere, una lettera che identifica l'etnia come Hispanic/Latin/Mexican/Black/Asian);
- eventuale arma usata;
- la latitudine e la longitudine del luogo dove si è verificato il crimine.

Simulare il conteggio dei crimini nel database InfluxDB suddivisi per distretto (con indicazione anche dell'area di appartenenza del distretto), per classificazione (crimini più gravi, crimini meno gravi), con e senza arma.

Durante la fase di caricamento, impostare regole di warning e alarm per supportare un sistema di allerta criminale automatizzato:

- numero di crimini giornaliero
- numero di crimini violenti
- percentuale di crimini con arma

Impostare le soglie di warning e alarm secondo un criterio 5%-10% (per esempio, se si supera il 5% del valore massimo giornaliero, si emette un warning, idem se si supera il 10% per attivare un allarme). I valori fuori scala (coordinate nulle, date future oltre Nov 2025, tipi di crimine non validi) vanno esclusi dal caricamento.

***2) Interrogazioni analitiche.*** Effettuare le seguenti analisi, utilizzando Python (`influxdb-client`, `pandas`, `matplotlib/plotly`) sulle collezioni create in precedenza:

- calcolare il numero medio giornalieri di crimini per area, raggruppato per stagione (autunno, inverno, primavera, estate) e visualizzare il risultato in un grafico a barre; quali aree mostrano un picco di criminalità stagionale?
- per un mese a scelta, calcolare la distribuzione delle categorie di crimine (parte 1 o 2) per diverse fasce orarie (diurna/notturna) e visualizzare il risultato in un grafico a barre; i crimini violenti sono maggiormente distribuiti di notte?
- identificare per ogni area la giornata con il maggior numero di crimini violenti e visualizzare il risultato un una tabella (area, data, conteggio); in quali mesi si concentrano i picchi più elevati?
- calcolare, per ogni categoria di crimine e per ogni anno, l'età media delle vittime.

Rieseguire tutte le query sopra filtrando per singola stagione (`season`). Misurare e confrontare i tempi di risposta per ciascun caso (query globali vs. stagionali).

***3) Analisi avanzata tramite clustering incrementale.*** Applicare un algoritmo di *clustering incrementale* sui pattern di criminalità, usando come variabili il numero di crimini, la percentuale di crimini con arma, la percentuale di crimini notturni, la percentuale di crimini violenti, aggregando per area, stagione, anno. L'obiettivo è individuare aree simili nei pattern di criminalità. Verificare se i cluster cambiano significativamente tra 2020 e 2025.

**Dettagli sulla consegna –** Predisporre un notebook Python (o un file .py) in cui è riportato lo svolgimento di tutti gli esercizi e un file PDF in cui sono riportati i commenti sui risultati laddove esplicitamente richiesto. Includere tutti i file in uno zip e caricare l'archivio su Moodle.