# Guida allo sviluppo del progetto (obbligatoria)

Eseguire un case study a scelta tra quelli elencati nel file use_cases.pdf

1) Implementare il case study scelto con due database a tua scelta tra Oracle, MongoDB, Cassandra, HBase e Neo4j.
2) Definire 4 dataset di dimensione crescente e inserirli nei due database scelti con lo stesso contenuto informativo (il risultato di una query deve essere lo stesso nei due database scelti). Nomina il 4 dataset 100%, 75%, 50% e 25% che indicano rispettivamente il dataset con il 100% del contenuto informativo, il set di dati con il 75% del contenuto informativo, ecc. Il dataset può essere scaricato dalla pagina del case study (se disponibile) o creato in modo casuale utilizzando un dato casuale generatore (es. Mockaroo, https://www.mockaroo.com/) o tramite linguaggio di programmazione (es. Faker libreria di Python, https://faker.readthedocs.io/en/master/).
3) Effettuare un confronto tra i due database con lo stesso dataset definendo 4 query con gradi crescenti di complessità dal punto di vista del numero dei soggetti coinvolti e del filtri di selezione.
4) Scegliere un linguaggio di programmazione a scelta per interagire con i due database effettuare operazioni di data entry (ove necessario) ed esecuzione di query.

Gli esperimenti devono essere automatizzati. Ogni esperimento (ad es. il test del tempo di esecuzione della query) deve essere eseguito almeno 30 volte considerando il valore medio e gli intervalli di confidenza al 95% (un tutorial su come tracciare gli intervalli di confidenza in un istogramma sono disponibili cliccando qui). I risultati devono essere salvati in un foglio elettronico da elaborare e rappresentare graficamente.

Gli istogrammi devono essere realizzati considerando i tempi di risposta espressi in millisecondi (msec). Per ogni query, tracciare i risultati man mano che le dimensioni del set di dati cambiano (25%, 50%, 75% e 100% del contenuto informativo).

Poiché molte soluzioni DBMS NoSQL utilizzano meccanismi di memorizzazione nella cache, con la stessa dimensione del set di dati, considera separatamente il tempo della prima esecuzione e il valore medio delle successive 30 esecuzioni per un totale di 31 prove. Per ogni

query stabilita devono essere creati due istogrammi: uno con la prima esecuzione volte al variare della dimensione del set di dati, uno con i tempi di esecuzione medi al variare della dimensione del dataset.

5) Dimostrare quale database a parità di condizioni hardware/software mostra un migliore tempo di esecuzione delle query .

6) Creare un breve report (max 15-20 pagine) che includa le seguenti sezioni:

- Problema affrontato (max 2 punti)
- Soluzione DBMS considerata (breve descrizione delle principali caratteristiche del database utilizzato) (max 2 punti)
- Design (contenente descrizione del modello dati utilizzato) (max 6 punti)
- Implementazione (contenente il codice utilizzato per l'inserimento dei dati e per l'implementazione di ciascuna interrogazione) (max 6 punti)
- Esperimenti (contenenti tabelle con i tempi di risposta ottenuti e relativi istogrammi) (max 6 punti)
- Conclusioni (max 3 punti)

Verranno assegnati da 0 a 6 punti aggiuntivi per la presentazione orale del progetto.

*Il giorno prima dell'esame lo studente deve inviare al docente via email:

1. il report sia in formato docx che pdf;
2. il foglio di calcolo comprensivo dei risultati sperimentali e degli istogrammi;
3. un collegamento a una cartella OneDrive che include il codice sviluppato e un dump del set di dati più breve (25% del contenuto informativo).

**Il giorno dell'esame gli studenti devono discutere anche con il docente il codice sviluppato utilizzando il proprio laptop o utilizzando un PC del laboratorio di informatica.