Basi di dati

Maurizio Lenzerini

Dipartimento di Ingegneria Informatica
Automatica e Gestionale "Antonio Ruberti"
Università di Roma "La Sapienza"

Anno Accademico 2023/2024

http://www.diag.uniroma1.it/~lenzerini/?q=node/44

1. Introduzione

1.1 introduzione al corso

1.2 introduzione alle basi di dati

Il corso di Basi di Dati

- È un corso di 6 crediti
- Laurea in Ingegneria Informatica e Automatica
- È previsto al terzo anno
- Collegato a questo corso c'è il corso di Data Management del Master of Science in Engineering in Computer Science (laurea magistrale), Master in Data Science and Master in AI and Robotics

Aspetti organizzativi del corso

Docente: Maurizio Lenzerini (http://www.diag.uniroma1.it/lenzerini)

Ricevimento:

- Martedì ore 17:00 on-line nella stanza virtuale all'indirizzo https://meet.google.com/hzy-save-oqw (controllare sempre il sito del docente per eventuali avvisi dell'ultimo momento)

Sito del corso di Basi di dati

http://www.diag.uniroma1.it/lenzerini/?q=node/44

Pagina MOODLE del corso di Basi di Dati

https://elearning.uniroma1.it/course/view.php?id=17068

Aspetti organizzativi del corso

Lezioni

- **Lunedì**, ore 09:00 – 11:00 (Sede Marco Polo, aula 204 + online)
- **Mercoledì**, ore 16:00 – 18:00 (Sede Marco Polo, aula 204 + online)
- **Giovedì**, ore 10:00 – 12:00 (Sede Marco Polo, aula 204 + online)

Esercitazioni

- in laboratorio il **venerdì** dalle 08:00 alle 11:00 in aula 16
- durante le lezioni

Lezioni ed esercitazioni vengono trasmesse on-line e registrate

L'esame viene superato con

- test di idoneità su SQL
- prova d'esame, a cui si partecipa solo dopo aver superato il test SQL, e che è composto da
  - prova scritta
  - prova orale (solo se necessaria)

Aspetti organizzativi del corso

- Materiale didattico
  - Slides delle lezioni (le slides di ogni parte del corso sono scaricabili dal sito del corso in MOODLE con qualche giorno di anticipo rispetto all'inizio della parte stessa)
  - Facoltativo:
    - Ramez A. Elmasri, Shamkant B. Navathe. Sistemi di basi di dati. Fondamenti e complementi. Pearson (collana: Informatica), 2017

    oppure

    - Atzeni, Ceri, Paraboschi, Torlone, Database Systems - Concepts, Languages and Architectures, McGraw-Hill
- Ulteriore materiale disponibile sulla pagina web
  - calendario e contenuto delle lezioni
  - testi e soluzioni di esercitazioni
  - documentazione sul DBMS adottato e altro materiale
  - esercizi di esame – testi e soluzioni (anni accademici precedenti)

I dati sono il nuovo petrolio

"Data is the new oil"
(**Clive Humby**, UK Mathematician, 2006)

- *Just like oil was a natural resource powering the last industrial revolution, data is going to be the natural resource for the new industrial revolution*

- *Oil is valuable, but if unrefined it cannot really be used. It has to be changed into gas, plastic, chemicals, etc. to create a valuable entity that drives profitable activity; so must data be prepared, broken down, and analyzed for it to have value*

Differenze tra dati e petrolio

- Al contrario del petrolio, i dati vengono generati ad un ritmo impressionante (nei prossimi due anni si stima verranno generati circa 40 zettabytes di dati, equivalenti a 4 milioni di anni di video HD, o a 5 miliardi di Libraries of Congress).
- È ormai facile produrre quantità enormi di dati (se il 25% delle auto di San Francisco fornissero i dati relativi ai loro percorsi, la quantità di dati che produrrebbero annualmente sarebbe la stessa generata da Twitter fino all'anno scorso).
- Il petrolio è un bene "single-use", mentre i dati possono essere riusati, condivisi, collegati per soddisfare nuove esigenze e per condurre nuove analisi.
- I dati devono essere organizzati e governati in modo razionale e memorizzati in forme usabili e accessibili (si calcola che solo il 10% dei dati siano organizzati in modo corretto e funzionale alle analisi e alle correlazioni che potrebbero valorizzarli). **Illustriamo, a questo proposito, un esempio paradigmatico sulla gestione dei dati relativi alla pandemia.**

COVID: il problema della Regione Lombardia

- **Da maggio 2020 a gennaio 2021** - in questo lasso di tempo la Lombardia segnala una grande quantità di casi, significativamente maggiore di quella osservata in altre regioni, ciascuno senza stato clicnico associato, ma con una specifica data di inizio sintomi. Basandosi su questa combinazione del valore dei campi "data inzio sintomi" e "stato clinico", tali casi sono considerati di carattere sintomatico.
- **29 maggio 2020 –** A partire da questa data la Regione Lombardia riceve settimanalmente il "Report di qualità e completezza dei dati" in cui vengono segnalate da ISS anomalie nei dati (valori incongruenti di data di inizio sintomi e stato clicnico).
- **7 gennaio 2021** - Gli epidemiologi dell'ISS segnalano quella che è a tutti gli effetti un'anomalia rispetto a tutte le altre Regioni e chiedono ai tecnici della Lombardia di verificare i loro dati, in particolare riguardo alla completezza dei campi relativi allo stato clinico. La richiesta viene fatta in ragione del fatto che alle regioni spetta il compito della verifica dei dati e sulla loro correttezza si basa l'attendibilità della stima dell'Rt elaborata dall'ISS.

COVID: il problema della Regione Lombardia

- **13 gennaio 2021** - Sulla base dei dati forniti, nella settimana del 13 gennaio 2021 viene attribuito alla Lombardia un Rt di 1,4 che manda in zona rossa la Regione.
- **19 gennaio 2021** - Nel corso di una riunione tecnica richiesta dalla Regione Lombardia viene segnalata l'ipotesi che la mancata compilazione della voce relativa allo stato clinico potrebbe essere alla base di un calcolo errato dell'Rt.
- **20 gennaio 2021 -** La Regione Lombardia invia come di consueto l'aggiornamento del suo database. In tale aggiornamento si realizza anche una rettifica dei dati pregressi. In particolare, viene cambiato il numero di casi in cui viene riportata una data inizio sintomi e, tra quelli con una data di inizio sintomi, si inserisce la compilazione del campo "stato clinico". Complessivamente, questi cambiamenti riducono in modo significativo il numero di casi che hanno i criteri per essere classificati come sintomatici e pertanto inclusi nel calcolo dell'Rt basato sulla data inizio sintomi dei soli casi sintomatici.

Obiettivi del corso

- *conoscenza delle basi di dati, dei principi della gestione dei dati e dei DBMS (sistemi di gestione di basi di dati) relazionali dal punto di vista degli utenti e degli sviluppatori di applicazioni*
- *conoscenza del modello relazionale*
- *conoscenza di SQL*
- *conoscenza del modello Entità-Relazione e capacità di applicazione di una metodologia di progettazione di basi di dati relazionali basata su tale modello*
- *introduzione ai modelli e ai sistemi di basi di dati NoSQL*

Obiettivi del corso della laurea magistrale "Data Management":

- *conoscenza di un Data Manager dal punto di vista di un amministratore di basi di dati e di un progettista di strumenti per la gestione dei dati*
- *conoscenza di problematiche avanzate di gestione di dati in applicazioni informatiche*

Programma di massima del corso

1. Introduzione
   - il concetto di basi di dati
   - introduzione ai sistemi di gestione
2. Il modello relazionale
   - basi di dati relazionali
   - algebra relazionale
3. Il linguaggio SQL
   - definizione ed utilizzo di una base di dati
   - utilizzo di una base di dati
4. Accesso alle basi di dati da software
   - principi di accesso ai dati da software
   - JDBC
5. La progettazione concettuale
   - modello entità-relazione
   - metodologia di progettazione concettuale
6. La progettazione logica-fisica
   - metodologia di progettazione logica
   - cenni alla progettazione fisica
7. Basi di dati NoSQ
   - Modelli NoSQL
   - Il sistema NoSQL MongoDB

1. Introduzione

1.1 introduzione al corso

1.2 introduzione alle basi di dati

Risorse di una organizzazione

- le risorse di una organizzazione:
  - persone
  - denaro
  - materiali
  - **dati e informazioni (sistema informativo)**

- funzioni di un sistema informativo
  - raccolta, acquisizione delle informazioni
  - archiviazione, conservazione delle informazioni
  - elaborazione delle informazioni
  - distribuzione, scambio di informazioni
  - il concetto di “sistema informativo” è indipendente da qualsiasi forma di automatizzazione

Base di dati

Due accezioni:

- **Collezione di dati in memoria secondaria**
- **Collezione di dati in memoria secondaria gestita da un apposito sistema software, chiamato DBMS** (Data Base Management Systems, o Sistema di Gestione di Basi di Dati).

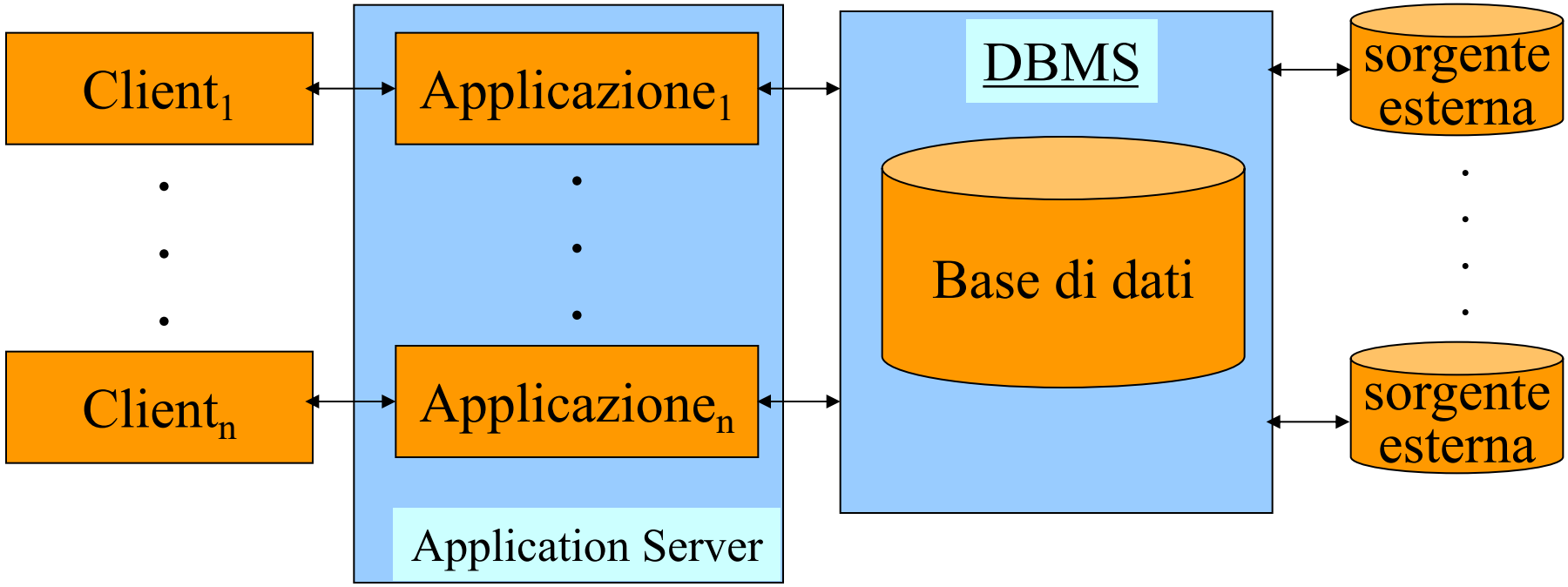
Architettura a tre livelli del software

Il DBMS è il sistema responsabile della gestione dei dati: i dati sono accessibili all'esterno solo tramite il DBMS

1 *Livello della presentazione*

2 *Livello della logica delle applicazioni*

3 *Livello dei dati*



Sistema di gestione di basi di dati
Data Base Management System — DBMS

- Sistema (**prodotto software**) in grado di gestire **collezioni di dati** che siano (anche):
  - **grandi** (di dimensioni molto maggiori della memoria centrale dei sistemi di calcolo utilizzati normalmente)
  - **persistenti** (con un periodo di vita indipendente dalle singole esecuzioni dei programmi che le utilizzano)
  - **condivise** (utilizzate da applicazioni diverse)

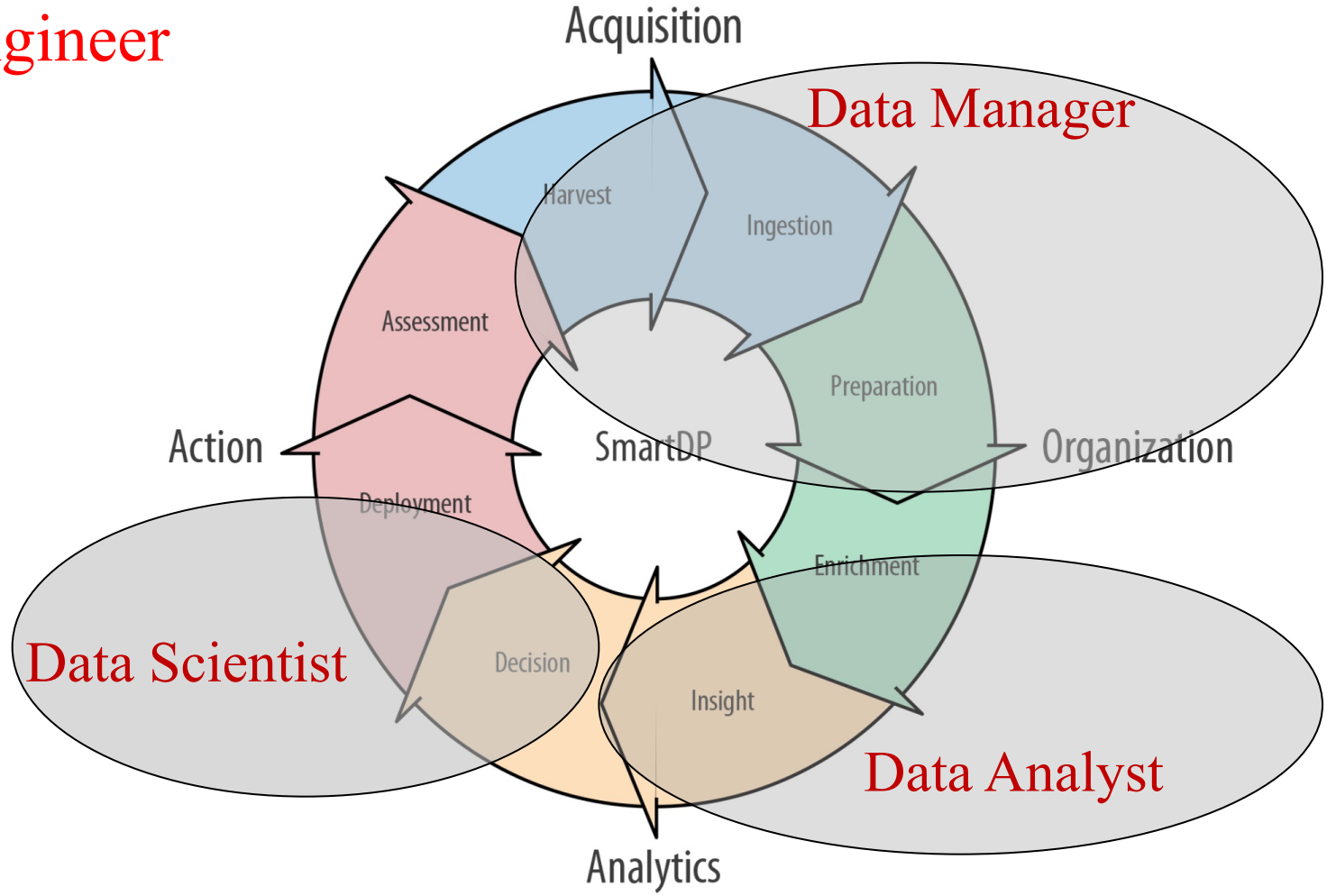
  garantendo:
  - **affidabilità** (resistenza a malfunzionamenti hardware e software)
  - **privatezza** (con una disciplina e un controllo degli accessi),
  - **efficienza** (utilizzare al meglio le risorse di spazio e tempo del sistema)
  - **efficacia** (rendere produttive le attività dei suoi utilizzatori).
- Ogni DBMS è basato su un modello dei dati, ovvero insieme di costrutti utilizzati per organizzare i dati di interesse e descriverne le operazioni. Noi ci riferiremo principalmente a DBMS basati sul **modello relazionale**, a sua volta basato sulla nozione di **relazione**.

Un po' di storia

- **Inizio anni '60**: Charles Bachman (General Eletric) progetta il primo DBMS (Integrated Data Store), basato sul modello reticolare. Bachman vincerà l'*ACM Turing Award* nel 1973.
- **Fine anni '60**: l'IBM sviluppa l'Information Management System (IMS), basato sul modello gerarchico e usato tutt'oggi.
- **1970**: Edgar Codd (IBM) propone il modello relazionale. Codd vincerà l'*ACM Turing Award* nel 1981.
- **Anni '80**: il modello relazionale vince sugli altri, e i DBMS basati su tale modello si diffondono. Il linguaggio SQL viene standardizzato come linguaggio per DBMS basati sul modello relazionale.
- **Anni '90**: sulla spinta di intense ricerche, i DBMS relazionali divengono sempre più sofisticati e diffusi (DB2, Oracle, Informix, ecc.). Nel 1999 James Gray (http://en.wikipedia.org/wiki/Jim_Gray_(computer_scientist)) vince l'*ACM Turing Award* per il suo contributo alla gestione delle transazioni.
- **Anni 2000**: i DBMS si integrano con il contesto generale dello sviluppo del software e con strumenti WEB
- **Anni 2010**: la gestione dei dati diventa sempre più pervasiva; nasce il concetto di "Big Data", secondo cui tutto il web si può vedere come una grande basi di dati, che necessità di tecniche e metodologie nuove
- **2014**: Michael Stonebraker vince l'*ACM Turing Award* "for fundamental contributions to the concepts and practices underlying modern database systems".

Il ciclo di gestione dei dati

Termine riassuntivo:
Data Engineer



Il ciclo di gestione dei dati

Termine riassuntivo:
Data Engineer

Questo corso: le basi per il