UNIVERSITÀ DEGLI STUDI DI PALERMO

DIPARTIMENTO	Matematica e Informatica
ANNO ACCADEMICO OFFERTA	2021/2022
ANNO ACCADEMICO EROGAZIONE	2021/2022
CORSO DILAUREA MAGISTRALE	INFORMATICA
INSEGNAMENTO	ELABORAZIONE DATI
CODICE INSEGNAMENTO	19742
MODULI	Si
NUMERO DI MODULI	2
SETTORI SCIENTIFICO-DISCIPLINARI	INF/01
DOCENTE RESPONSABILE	GIANCARLO RAFFAELE Professore Ordinario Univ. di PALERMO
ALTRI DOCENTI	GIANCARLO RAFFAELE Professore Ordinario Univ. di PALERMO ROMBO SIMONA ESTER Professore Ordinario Univ. di PALERMO
CFU	12
PROPEDEUTICITA'	
MUTUAZIONI	
ANNO DI CORSO	1
PERIODO DELLE LEZIONI	2° semestre
MODALITA' DI FREQUENZA	Facoltativa
TIPO DI VALUTAZIONE	Voto in trentesimi
ORARIO DI RICEVIMENTO DEGLI STUDENTI	**GIANCARLO RAFFAELE** Lunedì 15:00 17:00 Stanza 106 Dipartimento di Matematica ed Informatica Giovedì 15:00 17:00 Stanza 106 Dipartimento di Matematica ed Informatica **ROMBO SIMONA ESTER** Lunedì 9:30 13:30 In presenza: Via Archirafi 34, Piano II, Stanza 220 - Telematico: via Microsoft Teams o altro canale - In entrambi i casi, e' consigliabile prenotarsi tramite email alla docente

DOCENTE: Prof. RAFFAELE GIANCARLO

PREREQUISITI	Conoscenze di base su: algebra e geometria, algoritmi e strutture dati, progettazione di basi di dati. Conoscenze avanzate di programmazione.
RISULTATI DI APPRENDIMENTO ATTESI	Conoscenza e capacita' di comprensione Acquisizione di conoscenze sull'organizzazione dei sistemi informativi aziendali, sull'utilizzo e la progettazione dei sistemi di supporto alle decisioni, sulle problematiche correlate al big data management. Acquisizione degli strumenti avanzati per l' analisi ed il progetto di algoritmi. Capacita' di utilizzare il linguaggio specifico proprio di queste discipline specialistiche. Capacita' di applicare conoscenza e comprensione Capacita' di analizzare problematiche relative a contesti caratterizzati da grosse moli di dati, anche in ambito industriale, e proporre soluzioni progettuali per la gestione e analisi dei dati in tali contesti ed il supporto alle decisioni. Capacita' di sviluppare software basato su algoritmi efficienti per grosse quantita' di dati. Autonomia di giudizio Capacita' di analizzare e valutare soluzioni per la gestione di grosse quantita' di dati. Capacita' di progettare sistemi di supporto alle decisioni mediante l'analisi delle specifiche tecniche fornite. Essere in grado di valutare le implicazioni e i risultati degli studi algoritmici presentati durante il corso e della complessita' computazionale dei problemi ad essi associati. Abilita' comunicative Capacita' di descrivere soluzioni di progettazione di sistemi informativi complessi e sistemi di supporto alle decisioni e analizzarne le prestazioni. Capacita' di cooperare per determinare delle soluzioni di progettazione appropriate in contesti applicativi caratterizzati dalla presenza di grosse quantita' di dati. Capacita' di esporre i risultati degli studi algoritmici, anche ad un pubblico non esperto. Essere in grado di sostenere l'importanza ed evidenziare le ricadute tecnologiche delle teorie studiate. Capacita' d'apprendimento Capacita' di aggiornamento attraverso la consultazione di testi avanzati e di pubblicazioni scientifiche correlate agli argomenti affrontati durante il corso. Capacita' di seguire, utilizzando le conoscenze acquisite, sia master di secondo livello, che corsi di approfondimento e seminari specialistici nell'ambito della Business Intelligence e del progetto di algoritmi e strutture dati. Competenze acquisite Strumenti per l'analisi di problematiche relative a grosse quantità di dati anche in ambito industriale. Strumenti per analisi e progetto di algoritmi su grosse quantità di dati, sia in ambito scientifico che industriale.
VALUTAZIONE DELL'APPRENDIMENTO	Prova progettuale, prova individuale. Prova progettuale (per il modulo di Big Data): consistera' nella realizzazione e implementazione di un progetto basato su tecnologie di Big Data. Il progetto rifletterà il grado di complessità tipico dei contesti reali (es. Industrie e Grandi Organizzazioni) E' obbligatorio che il progetto sia realizzato da un team costituito da un minimo di due ad un massimo di cinque studenti, al fine di favorire lo sviluppo della capacita' di lavorare in team. Verranno effettuate due revisioni di progetto, durante le quali la docente verifichera' il contributo individuale di ciascun componente del team e assegnera' una valutazione alle diverse fasi di progettazione. Ciascun studente sarà chiamato ad esporre dinanzi ad un insieme di "peer", costituiti nel caso più semplice da suoi colleghi, la soluzione a cui ha contribuito. Alla fine al progetto verra' attribuito un punteggio complessivo, pesato per ciascuno studente in funzione del contributo individuale. Contribuiranno ad una valutazione positiva l'originalita' della soluzione proposta, il rigore metodologico, la dimostrazione di avere acquisito appieno gli strumenti tecnici forniti dal corso. I requisiti per conseguire il punteggio minimo per accedere alla seconda parte al colloquio orale per questo modulo consistono nella capacita' di realizzare un progetto che, almeno in modo basilare, rispetti le specifiche assegnate e sia corretto. Prova individuale: Consistera' nella risoluzione di uno specifico problema assegnato per ciascuno dei due moduli e in una prova orale. Quest'ultima servira' per verificare le conoscenze acquisite durante il corso, l'autonomia nell'approfondire contenuti anche complessi e la capacita' individuale di trovare soluzioni a problemi proposti. Per il modulo di Big Data Management, lo studente produrra' anche una breve tesina e presentazione in slide della stessa. Quindi, verra' verificata la capacita' critica e l'autonomia di giudizio dello

	studente attraverso un numero adeguato di domande. La valutazione della prova individuale sara' distinta per i due moduli e, per quello di Big Data Management, andra' a integrare il punteggio conseguito da ciascuno studente a seguito della valutazione del progetto. La valutazione finale sara' la media delle valutazioni per ciascuno dei due moduli. In particolare: 18-21: Conoscenza sufficiente di tutti gli argomenti del programma. 22-24: Conoscenza discreta di tutti gli argomenti del programma. 25-27: Conoscenza buona di tutti gli argomenti del programma. 28-30: Conoscenza ottima di tutti gli argomenti del programma. Lode. Viene attribuita nel caso in cui, se il punteggio e' pieno, lo studente si e' distinto in entrambi i moduli per la capacita' di applicare con grande autonomia i contenuti dell'insegnamento.
ORGANIZZAZIONE DELLA DIDATTICA	Lezioni frontali

MODULO
SCIENZA E INGEGNERIA DEGLI ALGORITMI

Prof. RAFFAELE GIANCARLO

TESTI CONSIGLIATI

William J.Cook, William H. Cunningham, William R. Pulleyblank, Alexander SchrijverI. Combinatorial Optimization, Wiley 199, ISBN: 978-0-471-55894-1

Per le parti che rigurdano algoritmi di approssimazione- For the part regarding approximation algorithms

Robert Endre Tarjan. Data Structure and Network Algoritms, SIAM 1984, ISBN 10: 0898711878

Per le parti che riguardano strutture dati in memoria interna-for the part regarding data structures in internal memory

Camil Demetrescu, Irene Finocchi,Giuseppe F. Italiano, Algoritmi e Strutture dati, McGraw Hill, 2005, ISBN-10 8838664684

Per le parti che riguardano l'analisi ammortizzata di algoritmi. For the part regarding amortised analysis of algorithms.

H. Cormen. C. Leiserson, R, Rivest, C. Stein Introduzione agli algoritmi e strutture dati, McGraw Hill, 2001, ISBN-10 883866515X
Per le parti che riguardano hashing e schemi di approssimazione-for the parts regarding hashing and approximation schemes

Materiale distribuito dal docente.
Tutto il resto del programma The remaining part of the syllabus.

TIPO DI ATTIVITA'	B
AMBITO	50341-Discipline Informatiche
NUMERO DI ORE RISERVATE ALLO STUDIO PERSONALE	102
NUMERO DI ORE RISERVATE ALLE ATTIVITA' DIDATTICHE ASSISTITE	48

OBIETTIVI FORMATIVI DEL MODULO

Esporre lo studente a tecniche avanzate di progetto ed analisi di algoritmi. In particolare, si copre tutto lo spettro delle strutture dati dinamiche e degli algoritmi, con approfondito studio di complessita' computazionale di problemi intrattabili o che coinvolgono grosse quantita' di dati e di loro soluzioni approssimate.

PROGRAMMA

ORE	Lezioni
4	Analisi ammortizzata: Metodo dei crediti. Metodo del potenziale. Analisi Sperimentale di Algoritmi
6	STRUTTURE DATI AVANZATE IN MEMORIA INTERNA: Bilanciamento Red Black Trees ed analisi delle operazioni su di essi. Linking and Cutting Trees ed analisi delle operazioni su di essi. Self-organizing List.
6	Self-adjusting binary trees ed analisi delle operazioni su di essi. Self-organizing Data Structure. Self-organizing List.
6	Dizionari Succinti in Spazio: Universal e Perfect Hashing.
5	Dizionari Succinti in Spazio: Bloom Filters.
6	MODELLO STREAMING. Motivazione. Il modello di calcolo streaming. Esempi di algoritmi streaming per mining di grosse quantita' di dati.
6	ALGORITMI STREAMING. Frequency Moments. Sketches.
4	ANALISI E SCHEMI DI COMPRESSIONE DATI Schemi di compressione, schemi di compressione adattivi. Ingegneria di compression boosting. Strutture dati efficienti per compressione dati. Benchmark per analisi di compressione.
3	TEORIA DEI PROBLEMI NP COMPLETI E APPROSSIMAZIONI POLINOMIALI Schemi di Approssimazione Polinomiale Inapprossimabilita' di Problemi
2	Metodi approssimati, Metodi euristici. TSP con disuguaglianza triangolare;

MODULO
BIG DATA MANAGEMENT

Prof.ssa SIMONA ESTER ROMBO

TESTI CONSIGLIATI

Parti su DATA WAREHOUSE: M. Golfarelli, S. Rizzi, “Data Warehouse – Teoria e pratica della progettazione”, Seconda Edizione, McGraw Hill, 2005, ISBN: 8838662916.
Parti su FRAMEWORK, TECNOLOGIE, BIG DATA MINING: Jure Leskovec, Anand Rajaraman, Jeff Ullman. "Mining of Massive Datasets", Third Edition, 2020 (disponibile online a http://www.mmds.org/).
Parti su DATA MINING: J. Han, M. Kamber, “Data Mining: Concepts and Techniques”, Morgan Kaufmann Publishers, 2006.

TIPO DI ATTIVITA'	B
AMBITO	50341-Discipline Informatiche
NUMERO DI ORE RISERVATE ALLO STUDIO PERSONALE	102
NUMERO DI ORE RISERVATE ALLE ATTIVITA' DIDATTICHE ASSISTITE	48

OBIETTIVI FORMATIVI DEL MODULO

Obiettivo del corso di Big Data Management e' fornire agli studenti di Informatica conoscenze sull'organizzazione dei sistemi informativi aziendali, sull'utilizzo e la progettazione dei sistemi di supporto alle decisioni, sulle problematiche correlate alla gestione di grosse quantita' di dati. Il corso iniziera' con una panoramica su Business Intelligence e si concentrera' su progettazione di Data Warehouse e OLAP. Successivamente, il corso affrontera' in modo approfondito la progettazione di database non relazionali, i framework di programmazione tipici del contesto Big Data (Map Reduce, Resilient Distributed Dataset, Tensor Flow, ecc.) e le relative tecnologie (Apache Hadoop, Apache Spark, ecc.). Si tratteranno problematiche emergenti, quali Data Lake e Extraction-Loading-Transformation (ELT). Si affronterà in modo dettagliato la risoluzione di problemi nell'ambito dell'Information Retrieval e Machine Learning nel distribuito, prendendo spunto dal mondo reale e apprendendo come apposite librerie di programmazione possano essere utilizzate a tal fine. Esempi di ciò sono il posizionamento delle pagine web, la pubblicità sul web, i sistemi di raccomandazione. Infine, si mostreranno degli scenari applicativi reali, provenienti dal mondo delle industrie e/o grandi organizzazioni.

PROGRAMMA

ORE	Lezioni
2	INTRODUZIONE AI BIG DATA Esigenze. Contesti applicativi. Una panoramica sulle tecnologie.
3	BUSINESS INTELLIGENCE E DATA WAREHOUSING Acquisizione dei dati. Elaborazione dei dati al fine di produrre informazioni. Archiviazione dei dati grezzi e delle informazioni. Trasmissione dei dati e delle informazioni. Presentazione dei dati e delle informazioni. Sistemi di supporto alle decisioni. Architetture per il Data Warehousing: architettura ad un livello, architettura a due livelli, architettura a tre livelli. I metadati. Qualita' di un Data Warehouse. Il modello multidimensionale e l’OLAP. Le principali operazioni OLAP. Modelli logici a supporto del Data Warehousing.
4	PROGETTAZIONE DI UN DATA WAREHOUSE Metodologia di progettazione di un Data Warehouse, scelta del processo, scelta della granularita, identificare e rendere conformi le dimensioni, scelta delle misure, memorizzare pre-calcoli nella tabella dei fatti, completare la tabella delle dimensioni, scelta della durata del database, tracciare le “slowly changing dimension”. Progettazione concettuale. Progettazione logica.
4	DATA CLEANING ETL. Problematiche e tecniche di data cleaning.
8	PROGETTAZIONE DI DATA WAREHOUSE REALI Studio dei moduli per l'integrazione e pulizia dei dati, OLAP. Loro applicazione per la risoluzione di un problema reale proveniente dal mondo industriale.
4	MEMORIZZAZIONE STRUTTURATA Database non relazionali. Tipologie di database NoSQL. Analisi di vantaggi e svantaggi dei database non relazionali. Esempi di implementazione di database NoSQL ed esercizi.
6	MAP REDUCE File System distribuiti. Paradigma MapReduce. Progettazione e implementazione di algoritmi in MapReduce (prodotti tra array, algebra relazionale, esercizi su conteggi di parole, PageRank, Google Ads, ecc.).
4	OTHER BIG DATA FRAMEWORKS AND TOOLS Resilient Distributed Datasets (RDD), TensorFlow, Apache Hadoop, Apache Spark.
5	BIG DATA MINING Richiami di Data Mining. Market Basket Analysis. Classificazione e Predizione. Alberi di decisione. K-Nearest Neighbor. Clustering. Outlier detection. Problematiche di data mining su grandi moli di dati, sfide e soluzioni. Intelligenza Artificiale e Big Data. Relazioni tra Data Mining, Machine Learning e Statistica. Reti sociali e Web advertising. Recommendation Systems. Studio della libreria Mllib di Spark e sua applicazione per Collaborative Filtering.
3	BIG NETWORKS Reti sociali. Reti biologiche. Clustering di reti. Validazione dei risultati di classificazione. Libreria GraphX di Spark.
5	BIG SEQUENCES E ALTRI PROBLEMI. Analisi di dati NGS e sequenze biologiche. Indicizzazione e compressione in MapReduce. Medicina personalizzata. Datasets e Data Frames. Data Lake. ELT vs ETL. Applicazioni dal mondo reale (e.g., Industria 4.0, supporto alle decisioni nell'ambito della medicina di precisione).