Principali informazioni sull'insegnamento

Denominazione dell'insegnamento	**Gestione e Analisi di Big Data**
Corso di studio	Laurea Magistrale in Data Science
Anno Accademico	2024/25
Crediti formativi universitari (CFU) / European Credit Transfer and Accumulation System (ECTS)	6 CFU
Settore Scientifico Disciplinare	INFO-01/A (ex INF-01) - Informatica
Lingua di erogazione	Italiano
Anno di corso	Secondo
Periodo di erogazione	1^ semestre, le date esatte sono riportate nel manifesto/regolamento
Obbligo di frequenza	La frequenza è fortemente raccomandata
Sito web del corso di studio	https://www.uniba.it/it/corsi/cdl-data-science/corso-di-laurea-in-data-science

Docente/i

Nome e cognome	Gianvito Pio
Indirizzo mail	gianvito.pio@uniba.it
Telefono	+39 0805442283
Sede	Dipartimento di Informatica, Via Orabona 4, 70125, Bari. Stanza n.525, 5^ piano.
Sede virtuale	Piattaforma: https://elearning.uniba.it/
Sito web del docente	https://gianvitopio.di.uniba.it/
Ricevimento (giorni, orari e modalità, es. su appuntamento)	Mercoledì dalle 14:00 alle 16:00, previo appuntamento

Syllabus	
Obiettivi formativi	Acquisizione di capacità di gestione e analisi di dati complessi, in termini di volume, di eterogeneità, di veridicità e di velocità di generazione. Acquisizione di capacità di progettazione di datawarehouse e di basi di dati basate su modelli NoSQL. Acquisizione di competenze relative alla progettazione e implementazione di algoritmi in grado di analizzare grandi moli di dati in ambienti distribuiti.
Prerequisiti	Dall'insegnamento di Gestione di dati strutturati e non strutturati: principi relativi alle basi di dati. Dall'insegnamento di Data Mining: metodi principali per l'analisi dei dati.

Contenuti di insegnamento (Programma)	1) Business Intelligence & Datawarehouse:
	• Introduzione agli obiettivi della business intelligence
	• Caratteristiche dei datawarehouse
	• OLTP vs. OLAP
	• Architettura dei datawarehouse
	• Il modello multidimensionale
	• Modelli logici per i datawarehouse
	• Progettazione di un datawarehouse
	2) Big Data: introduzione e storage
	• Introduzione
	• Definizioni
	• Caratteristiche e sfide dei big data
	• Metodologie di analisi dei big data
	• Memorizzazione dei dati tramite sistemi NoSQL
	  ○ Concetti preliminari
	  ○ Rilassamento delle garanzie di consistenza
	  ○ Tipi di sistemi NoSQL
	• Memorizzazione dei dati in database decentralizzati: la Blockchain
	  ○ Concetti preliminari
	  ○ Protocolli di consenso: il Proof of Work
	  ○ Introduzione agli Smart Contract in Ethereum
	3) Big Data: analisi
	• Il paradigma di programmazione MapReduce
	• Il framework Apache Spark
	• Analisi dei dati in Apache Spark
	Esercitazioni e laboratorio:
	• DBMS NoSQL MongoDB
	• Smart contract in Ethereum
	• Algoritmi distribuiti in Apache Spark

Testi di riferimento	Viktor Mayer-Schonberger, Kenneth Cukier. Big Data: A Revolution That Will Transform How We Live, Work, and Think, John Murray, 2013 Gli studenti che lo desiderano possono ottenere i testi in prestito dalla Biblioteca. Può convenire verificarne la disponibilità mediante il Sistema Bibliotecario di Ateneo https://opac.uniba.it/easyweb/w8018/index.php? e contattare la biblioteca per concordare il prestito.
Note ai testi di riferimento	I libri di testo sono integrati con le slide fornite dal docente ed eventuale materiale di approfondimento, resi disponibili sulla piattaforma (vedi sopra "sede virtuale"). Sulla medesima piattaforma sono anche disponibili esempi di tracce di prove scritte e di laboratorio, alcune corredate da esempi di svolgimento.

Organizzazione della didattica

Ore

Totali	Didattica frontale	Pratica (laboratorio, progetto, esercitazione, altro)	Studio individuale
150 ore	32 ore	30 ore	88 ore

CFU/ETCS

6 CFU	4 CFU	2 CFU	

Metodi didattici

Le lezioni frontali saranno dedicate all'apprendimento dei modelli teorici e dei concetti di base, coadiuvati da alcuni esempi. Le ore di esercitazione saranno dedicate sia all'esecuzione di esercizi in classe, anche coinvolgendo direttamente gli studenti nella risoluzione degli stessi, sia alla implementazione di datawarehouse e algoritmi distribuiti. Si prevede l'utilizzo della piattaforma di e-learning del dipartimento per la pubblicazione del materiale didattico, la discussione degli argomenti delle lezioni tra docente/studente e studenti/studenti, la condivisione dei risultati di laboratorio, la condivisione degli esercizi e la pubblicazione di materiale integrativo e di approfondimento.

Risultati di apprendimento previsti

Conoscenza e capacità di comprensione	Il corso si propone di introdurre il discente alle tematiche della gestione di grandi moli di dati e alla loro analisi attraverso algoritmi distribuiti. Per la gestione saranno studiati modelli di memorizzazione basati su datawarehouse e database NoSQL, mentre per l'analisi distribuita sarà introdotto il paradigma di programmazione MapReduce, adottato dal framework Apache Spark.
Conoscenza e capacità di comprensione applicate	Il discente sarà in grado di comprendere i limiti delle tecnologie tradizionali e di applicare paradigmi all'avanguardia volti a superarli. Tali paradigmi, in particolare, riguardano l'analisi di grandi moli di dati, slegandosi dal paradigma SQL classico e dalla restrizione all'uso di una singola macchina di calcolo. Queste competenze sono trasferite attraverso lezioni teoriche ed esercitazioni pratiche.
Competenze trasversali	**Autonomia di giudizio** Maturare capacità di giudizio e di prendere decisioni ponderate è esattamente lo scopo della progettazione di un'applicazione di Big Data Analytics. Pertanto, l'autonomia di giudizio è maturata durante l'applicazione pragmatica di scelte progettuali e l'analisi dei risultati ottenuti.

	Abilità comunicative Analogamente, per rendere fruibile, anche ai non esperti, la conoscenza estratta da una grande mole di dati, il discente deve apprendere a interpretarla, formalizzarla e presentarla nella maniera più chiara e adeguata possibile. Questo è un passaggio fondamentale di un processo di Big Data Analytics, come, peraltro, di un processo di KDD. **Capacità di apprendere in modo autonomo** Il discente apprenderà concetti teorici e pratici che lo metteranno nella posizione di comprendere e utilizzare strumenti utili all'estrazione di conoscenza da grandi moli di dati, anche differenti, in maniera autonoma.

Valutazione	
Modalità di verifica dell'apprendimento	L'esame consiste in una prova scritta e nella discussione di un caso di studio. La prova scritta è costituita da domande aperte che possono riguardare sia argomenti di natura teorica che lo sviluppo di una soluzione a problemi analoghi a quelli trattati durante il corso.
Criteri di valutazione	Si richiede che lo studente sia in grado di individuare scenari tipici dei Big Data e affrontare le relative problematiche, in termini di memorizzazione e analisi degli stessi. Lo studente deve essere in grado di individuare le soluzioni tecniche più appropriate, tra quelle studiate. Sul piano pratico, lo studente dovrà dimostrare di saper progettare e implementare datawarehouse, progettare un database seguendo modelli NoSQL, e progettare e implementare algoritmi distribuiti in Apache Spark.
Criteri di misurazione dell'apprendimento e di attribuzione del voto finale	Il voto finale è attribuito in trentesimi. L'esame si intende superato quando il voto finale è maggiore o uguale a 18. L'accesso alla discussione del caso di studio richiede il superamento della prova scritta con un voto maggiore o uguale a 18.
Altro	Si suggerisce agli studenti di affidarsi esclusivamente alle informazioni/comunicazioni fornite sui siti ufficiali del Dipartimento di Informatica, ovvero sui gruppi social solo se costituiti e amministrati esclusivamente dai docenti dei relativi insegnamenti: ● https://www.uniba.it/it/ricerca/dipartimenti/informatica/didattica/corsi-di-laurea/corsi-di-laurea ● https://www.uniba.it/it/ricerca/dipartimenti/informatica ● https://elearning.uniba.it/ Le informazioni che tutti gli studenti dovrebbero conoscere sono scritte nei Regolamenti didattici e manifesti degli studi disponibili nel sito: ● https://www.uniba.it/it/ricerca/dipartimenti/informatica/didattica/corsi-di-laurea/corsi-di-laurea Si suggerisce agli studenti di diffidare delle informazioni e dei materiali circolanti su siti o gruppi social non ufficiali, poiché spesso sono risultati non affidabili, non corretti o incompleti. Per ogni dubbio, chiedere un incontro al docente secondo le modalità previste per il ricevimento.

	- Link al corso sulla piattaforma e-learning: https://elearning.uniba.it/course/view.php?id=4973 - Canale Telegram: https://t.me/lmds_bigdata_2023

Main information on the course

Course name	**Big Data Management and Analytics**	
Degree	Master's course in Data Science	
Academic year	2024/25	
European Credit Transfer and Accumulation System (ECTS), in Italian Crediti Formativi Universitari (CFU)		6 CFU
Settore Scientifico Disciplinare	INFO-01/A (ex INF-01) – Informatica	
Course language	Italian	
Year	Second	
Period	First semester, the exact dates are reported in the manifest	
Attendance obligation	None, but it is highly recommended to attend classes	
Web site of the degree	https://www.uniba.it/it/corsi/cdl-data-science/corso-di-laurea-in-data-science	

Teacher(s)

Name and Surname	Gianvito PIO
email	gianvito.pio@uniba.it
phone	+39 080 5442283
office	Dept. of Computer Science, Via Orabona 4, 70125, Bari. Room n.525, 5^ floor
e-learning platform	https://elearning.uniba.it/
Teacher's homepage	https://gianvitopio.di.uniba.it/
Office hours	Wednesday 14:00-16:00. Students should send an email to the teacher to require an appointment.

Syllabus

Course goals	Acquisition of management and analytical skills related to complex data, in terms of volume, heterogeneity, truthfulness and speed of generation. Acquisition of datawarehouse and database design skills based on NoSQL models. Acquisition of skills related to the design and implementation of algorithms capable of analysing large amounts of data in distributed environments.
Prerequisites/requirements	From the course of "Management of structured and unstructured data": basics about data bases. From the course "Data Mining": main methods for data analysis.
Course program	1) Business Intelligence & Datawarehouse: • Introduction to the objectives of business intelligence • Characteristics of datawarehouses • OLTP vs. OLAP • Datawarehouse architectures • The multidimensional model • Logical models for datawarehouses • Design of a datawarehouse

	2) Big Data: introduction and storage
	• Introduction
	• Definitions
	• Characteristics and challenges of big data
	• Big data analytic methodologies
	• Data storage with NoSQL systems
	  ○ Preliminary concepts
	  ○ Relaxation of consistency guarantees
	  ○ Types of NoSQL systems
	• Data storage in decentralized databases: the Blockchain
	  ○ Preliminary concepts
	  ○ Consensus protocols: Proof of Work
	  ○ Introduction to Smart Contracts in in Ethereum
	
	3) Big Data: analytics
	• The MapReduce programming paradigm
	• The Apache Spark framework
	• Data analysis in Apache Spark
	Exercises and laboratory:
	• MongoDB NoSQL DBMS
	• Development of Smart Contracts in Ethereum
	Distributed algorithms with Apache Spark

Books of reference	Viktor Mayer-Schonberger, Kenneth Cukier. Big Data: A Revolution That Will Transform How We Live, Work, and Think, John Murray, 2013 Students can also rent the books from the University Library. It may be useful to check the availability through the Sistema Bibliotecario di Ateneo https://opac.uniba.it/easyweb/w8018/index.php? and to contact the library to plan the book rent.
Notes to the books	Books are also integrated with the slides provided by the teacher, as well as by other material, make available on the platform (see “e-learning platform”). On the same platform, there will be also some examples of exams for both theory and laboratory, often together with some examples of solutions.

Organization of the didactic activities

Hours			
Total	Lectures	Practice sessions	Individual study
150 hours	32 hours	30 hours	88 hours
CFU/ETCS			
6 CFU	4 CFU	2 CFU	

Teaching methods

	The lectures will be dedicated to learning theoretical models and basic concepts, supported by some examples. The practice sessions will be dedicated both to the execution of exercises in the classroom, also directly involving the students in solving them, and to the implementation of datawarehouses and distributed algorithms. The e-learning platform of the department and/or the Microsoft Teams channel will be used for the publication of teaching material, the discussion of the topics of the lessons between teacher and students, the sharing of laboratory results, the sharing of exercises, and the publication of supplementary material.

Expected learning outcomes	
Knowledge and understanding	The course aims to introduce the learner to the issues of managing large amounts of data and their analysis through distributed algorithms. Programming models and NoSQL databases will be studied for the management, while the MapReduce programming paradigm, from the Apache Spark framework, will be used for the analysis.
Applying knowledge and understanding	The student will be able to understand the limits of traditional technologies and apply cutting-edge paradigms aimed at overcoming them. These paradigms, in particular, concern the analysis of large amounts of data, breaking away from the classic SQL paradigm and the restriction on the use of a single computing machine. These skills are transferred through theoretical lessons and practical exercises.
Other skills	• *Autonomy of judgment* Developing the ability to judge and make informed decisions is exactly the purpose of designing a Big Data Analytics application. Therefore, the autonomy of judgment is gained during the pragmatic application of design choices and the analysis of the results obtained. • *Communication skills* Similarly, to make knowledge extracted from a large amount of data accessible, even to non-experts, the student must learn to interpret, formalize and present data in the clearest and most appropriate way. This is a fundamental step of a Big Data Analytics process, as well as a KDD process. • *Ability to learn* The student will learn theoretical and practical concepts that will enable her/him to understand and apply useful tools for extracting knowledge from large amounts of data.

Assessment	
Assessment methods	The exam consists of a written test and the discussion of a case study. The written consists of questions that can present both theoretical topics and the development of a solution to problems similar to those seen during the course.
Evaluation criteria	The student is required to identify and address typical Big Data related problems. The student must be able to identify the most appropriate technical solutions, among those studied. On a practical level, the student will have to know how to design and implement a datawarehouse, design a database following NoSQL models, and design and implement distributed algorithms in Apache Spark.
Criteria for assessment and attribution of the final mark	The final grade is awarded out of 30. The exam is passed when the final grade is greater than or equal to 18. Access to the discussion of the case study requires passing the written test with a grade greater than or equal to 18.
Further information	The students should exclusively consider the information/communication provided on the official website of the Department of Computer Science, or on the social groups exclusively administrated by the teachers of the course. ● https://www.uniba.it/it/ricerca/dipartimenti/informatica/didattica/corsi-di-laurea/corsi-di-laurea ● https://www.uniba.it/it/ricerca/dipartimenti/informatica ● https://elearning.uniba.it/ The information that all students should know are reported in the regulations and manifests available at: ● https://www.uniba.it/it/ricerca/dipartimenti/informatica/didattica/corsi-di-laurea/corsi-di-laurea

	It is strongly suggested to not consider information or material made available on unofficial websites and social medias, because they are often unreliable, incorrect and incomplete. For any doubts, the students should ask for a meeting with the teacher. - *Link to the course on the e-learning platform: https://elearning.uniba.it/course/view.php?id=4973* - *Telegram Channel: https://t.me/lmds_bigdata_2023*