Principali informazioni sull'insegnamento	
Denominazione dell'insegnamento	Gestione e Analisi di Big Data
Corso di studio	Laurea Magistrale in Data Science
Anno di corso	II
Crediti formativi universitari (CFU) / European Credit Transfer and Accumulation System (ECTS):	6
SSD	INF-01 - Informatica
Lingua di erogazione	Italiano
Periodo di erogazione	Primo semestre
Obbligo di frequenza	-

Docente	
Nome e cognome	Gianvito Pio
Indirizzo mail	gianvito.pio@uniba.it
Telefono	+39 0805442203
Sede	Dip. Informatica - 5° Piano, Stanza 571
Sede virtuale	-
Ricevimento (giorni, orari e modalità)	Mercoledì dalle 11:30 alle 13:30, previo appuntamento via e-mail

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	Principi relativi alle basi di dati e ai metodi di 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 • Back, D. W., Goodman, N., & Hyde, J. (2013). Mondrian in Action: Open source business analytics. Manning Publications • Chodorow, K. (2013). MongoDB: The definitive guide. Powerful and scalable data storage. O'Reilly Media, Inc. • Bill Chambers, Matei Zaharia, Spark: The Definitive Guide: Big Data Processing Made Simple. O'Reilly & Associates Inc, 2018
Note ai testi di riferimento	I libri di testo sono integrati con le slide e le dispense del docente.

Organizzazione della didattica			
Ore			
Totali	Didattica frontale	Pratica (laboratorio, campo, esercitazione, altro)	Studio individuale
150	32	30	88
CFU/ETCS			
6	4	2	

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 e/o del canale Microsoft Teams 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* 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.

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 allo studente, durante le ore di studio individuale, di arricchire la propria conoscenza con un lavoro di approfondimento autonomo sulle ricche funzionalità messe a disposizione dalle tecnologie spiegate.

General information	
Academic subject	Big Data Management and Analytics
Degree course	M.Sc. in Data Science
Academic Year	2nd
European Credit Transfer and Accumulation System (ECTS)	6
Language	Italian
Academic calendar (starting and ending date)	1st semester
Attendance	-

Professor/ Lecturer	
Name and Surname	Gianvito Pio
E-mail	gianvito.pio@uniba.it
Telephone	+39 0805442203
Department and address	Computer Science Department – 5th Floor, Room 571
Virtual headquarters	-
Tutoring (time and day)	Wednesday from 11:30 to 13:30, by prior agreement via e-mail

Syllabus	
Learning Objectives	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.
Course prerequisites	Principles related to databases and methods of data analysis.

Contents	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 and bibliography	• Viktor Mayer-Schonberger, Kenneth Cukier. Big Data: A Revolution That Will Transform How We Live, Work, and Think, John Murray, 2013 • Back, D. W., Goodman, N., & Hyde, J. (2013). Mondrian in Action: Open source business analytics. Manning Publications • Chodorow, K. (2013). MongoDB: The definitive guide. Powerful and scalable data storage. O'Reilly Media, Inc. • Bill Chambers, Matei Zaharia, Spark: The Definitive Guide: Big Data Processing Made Simple. O'Reilly & Associates Inc, 2018
Additional materials	Textbooks are integrated with the teacher's slides and handouts.

Work schedule			
Total	Lectures	Hands on (Laboratory, working groups, seminars, field trips)	Out-of-class study hours/ Self-study hours
Hours			
150	32	30	88
ECTS			
6	4	2	

Teaching strategy	
	The lectures will be dedicated to learning theoretical models and basic concepts, supported by some examples. The hours of practice 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 on:	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 on:	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.
Soft 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 and feedback	
Methods of assessment	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.
Additional information	
	During the hours of individual study, students are advised to enrich their knowledge with an independent study of the rich functionalities made available by the technologies explained.