Principali informazioni sull'insegnamento	**A.A. 2020-2021**
Titolo insegnamento	Gestione e Analisi di Big Data
Corso di studio	Laurea Magistrale in Data Science
Crediti formativi	6
Denominazione inglese	Big Data Management and Analytics
Obbligo di frequenza	-
Lingua di erogazione	Italiano

Docente responsabile	Nome Cognome	Indirizzo Mail
	Gianvito Pio	gianvito.pio@uniba.it
	Gennaro Vessio	gennaro.vessio@uniba.it
Luogo e Orario di Ricevimento	Dip. Informatica 4° Piano, Lab CILAB/KDDE	Mercoledì dalle 11:30 alle 13:30 previo appuntamento via e-mail

Dettaglio credi formativi	Ambito disciplinare	SSD	Crediti
	INFORMATICO	INF-01 - Informatica	6 (4+2)

Modalità di erogazione	
Periodo di erogazione	Primo semestre
Anno di corso	Secondo anno
Modalità di erogazione	Lezioni frontali (4 CFU) ed esercitazioni guidate (2CFU), pari a un totale di 62 ore di lezioni frontali (32 ore di lezioni teoriche e 30 ore di esercitazioni guidate)

Organizzazione della didattica	
Ore totali	150
Ore di corso	62
Ore di studio individuale	88

Calendario	
Inizio attività didattiche	5 ottobre 2020
Fine attività didattiche	13 gennaio 2021

Syllabus	
Prerequisiti	Principi relativi alle basi di dati e ai metodi di analisi dei dati.
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. • *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.
Contenuti di insegnamento	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 o Concetti preliminari o Rilassamento delle garanzie di consistenza o Tipi di sistemi NoSQL 3) Big Data: analisi

	• Il paradigma di programmazione MapReduce • Il framework Apache Spark • Analisi dei dati in Apache Spark Esercitazioni e laboratorio: • Implementazione di Datawarehouse tramite Mondrian • DBMS NoSQL MongoDB • Algoritmi distribuiti in Apache Spark

Programma	
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.
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.
Metodi di valutazione	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.
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.