MODELLO D (inglese)			
General Information			
Academic subject	Big Data Analytics		
Degree course	Laurea Magistrale in Informatica - Computer Science (second-level degree in Computer Science)		
Curriculum	INGEGNERIA DELLA CONOSCENZA E INTELLIGENZA DELLE MACCHINE (KNOWLEDGE ENGINEERING AND MACHINE INTELLIGENCE)		
ECTS credits	6		
Compulsory attendance	no		
Language	Italiano		
Subject teacher	Name Surname	Mail address	SSD
	Michelangelo Ceci	michelangelo.ceci@uniba.it	ING-INF/05
ECTS credits details			
Basic teaching activities	Databases	ING-INF/05	6
Class schedule			
Period	First semester		
Year	2017/2018		
Type of class	Lectures Lab		
Time management			
Hours	47		
Hours of lectures	32 (4 credits)		
Tutorials and lab	15 (1 credits)		
Project	(1 credits)		
Academic calendar			
Class begins	25/09/2017		
Class ends	12/01/2018		
Syllabus			
Prerequisites/requirements	Database Systems (or Basi di Dati II)		
Expected learning outcomes (according to Dublin Descriptors) (it is recommended that they are congruent with the learning outcomes contained in the Didactic Regulation and Prospectus a.a. 2017-2018)	*Knowledge and understanding* Data analysis has replaced data acquisition as the bottleneck to evidence-based decision making --- we are drowning in it. Extracting knowledge from large, heterogeneous, and noisy datasets requires not only powerful computing resources, but also methodological basis and the appropriate programming abstractions. The abstractions that emerged in the last decade blend ideas from parallel databases, distributed systems, and programming languages to create a new class of scalable data analytics platforms that form the foundation for data science at realistic scales. In this course, you will learn the landscape of relevant systems, the principles on which they rely, their tradeoffs. You will learn how practical systems were derived from the frontier of research in computer science and what systems are coming on the horizon. NoSQL databases, MapReduce and the ecosystem it spawned and Spark architecture will be covered.		

	Applying knowledge and understanding The course will have a practical part in Lab and the students will be asked to design and implement a Big Data Analytics tool as a project. *Making informed judgements and choices* Making informed judgements and choices is exhactly the purpose of application design. In this course, the students will learn to autonomously design and implement a Big Data Analytics tool and perform analyses. *Communicating knowledge and understanding* In order to make the extracted knowledge actionable, the result of the analysis must be adequately presented. This is a fundamental step of the KDD process and, consequently this is a undamental step of Big Data Analytics. *Capacities to continue learning* The student will learn basic concepts that will make her/him on the position of use, understand and deploy any data mining method which extracts knowledge from large volume of data.
Contents	1) Big Data: introduction • Big Data: Definitions, Why and Where • Characteristics of Big Data and Dimensions of Scalability • Getting Value out of Big Data 2) Big Data: Data modelling and storage NoSQL systems are purely about scale rather than analytics, and are arguably less relevant for the practicing data scientist. However, they occupy an important place in many practical big data platform architectures, and data scientists need to understand their limitations and strengths to use them effectively. Introduction to Big Data Modeling and Management • NoSQL Concepts: Relaxing Consistency Guarantees • Types of NoSQL Systems • NoSQL: Systems 3) Big data: Analysis A tour through the important methods, algorithms, and techniques in data mining. • Knowledge Discovery from Databases: definition. the process of Knowledge Discovery from Databases. CRISP-DM: business understanding, data understanding, data preparation, • Modelling, evaluation, deployment. • Main methods that perform well on a variety of tasks, such as: Classification, Parametric and non-parametric Regression and Variable associations. 4) Big data: Programming

	• The MapReduce programming model, as a simplifying abstraction for parallel manipulation of massive datasets. • Systems: Getting Started with Hadoop MapReduce and Spark. • Big Data Analytics using Spark LAB: Spark-based solutions for data mining problems + NoSQL DBMS
Course program	
Bibliography	• Viktor Mayer-Schonberger, Kenneth Cukier. Big Data: A Revolution That Will Transform How We Live, Work, and Think, John Murray, 2013 • T. Mitchell Machine Learning, Morgan Kaufmann, 1997 • Richard J. Roiger, Michael W. Geatz. Introduction to Data Mining McGraw-Hill, 2003 • A. Azzalini, B. Scarpa Analisi dei dati e data mining, Springer, 2004 Slides presented during classes
Notes	
Teaching methods	Lectures, lab. All with the support of Slides prepared by the teacher.
Assessment methods (indicate at least the type written, oral, other)	The exam consists of an oral part and a laboratory part. The oral part aims at verifying the acquisition of all the topics addressed during lectures. The laboratory part is an oral discussion of a small project and aims to verify the capabilities of designing a complete analysis and present results. All using the technologies used during the laboratory. Between the parts there is no propedeuticity (i.e., they are independent each other). The mark obtained for both parts expires in May 2018. During the teaching period, two partial evaluations are planned. If successful, they substitute the oral exam.
Evaluation criteria (Explain for each expected learning outcome what a student has to know, or is able to do, and how many levels of achievement there are).	The student should prove to know all the concepts discussed during classes, as well as, show that (s)he is able to design and implement a system according to the best practices discussed.
Further information	