UNICUSANO
ROME

Code: ING-INF/05 **Credits: 9**
Matter: Algorithms and Data Structures for Big Data
Main language of instruction: Italian
Other language of instruction: English

Teaching Staff

Head instructor
Prof.ssa Paola Vocca - paola.vocca@unicusano.it

Introduction

1. *Objective of the course:*

The Course of Algorithms and Data Structures for Big Data aims to give the student a good knowledge of the principles that govern the design of systems for the management of large amount of data and allow them to model their behavior. The course proposes the basic concepts of distributed architectures for processing and storing Big Data, the study of algorithms and, more generally, of the techniques of pre-processing, dimensional reduction, clustering, classification, and prediction. In addition, the educational objective of the course is to provide the student with a detailed knowledge of the approaches to the storage and structuring of data, both relational and non-relational. The Etivity associated with the course develop the skills necessary to design and analyze systems and algorithms for Big Data.

Objectives

2. *Course Structure:*

The course is composed by audio-video prerecorded lessons that make up, together with slides and handouts, the study materials available on the platform.
There are also proposed self-assessment tests, asynchronous, that accompany the pre-recorded lessons and allow students to ascertain both the understanding and the degree of knowledge acquired of the contents of each of the lessons.
The interactive teaching is carried out in the forum of the "virtual class" and includes 4 Etivity that apply the knowledge acquired in theory lessons to the analysis and design of applications based on Big Data using environments for programming oriented to data analytics.

Competencies:

Knowledge and understanding.

The student at the end of the course will have knowledge of the computation models, architectures and infrastructures necessary for the processing of large quantities of data. The student will also be familiar with the problems related to the analysis of large quantities of data. Additionally, through the Etivity students will gain the ability to analyze infrastructure and systems for Big Data.

Application of knowledge.

The student will be able to analyze scenarios characterized by the presence of large amounts of data; you will also be able to provide appropriate design solutions for the construction of systems capable of managing such data, and to design efficient software systems for processing large amounts of data.

The Etivities provide for the application of theoretical knowledge to the design and implementation of Big Data-based applications within data analytics-oriented programming environments.

Making judgements.

The student will be able to evaluate the correctness of the different algorithmic and architectural solutions for the management of large amounts of data and will also be able to evaluate the performance of the different approaches by interpreting appropriate indicators. Finally, the student will be able to carry out bibliographic searches, to analyze and interpret the relevant sources, in order to analyze new paradigms, approaches, architectures and algorithms for the processing of large amounts of data.

Communication skills.

The student will be able to describe and hold conversations on issues relating to the design and management of systems for the management of large amounts of data, and to the resolution of typical problems of such systems, using adequate terminology.

Learning skills.

At the end of the course, the student will have knowledge of the fundamental notions necessary for the analysis and design of various applications that require the use of Big Data. This will allow him to identify the tools necessary to independently learn the operating principles of new tools for managing big data.

Syllabus

3. *Programme of the course:*

Subject 1.

Introduction: where the following topics are addressed: Introduction to the basic concepts of Big Data: terminology, main aspects, and examples of applications.

Subject 2.
Infrastructures the following topics are addressed: Infrastructure for the management of Big Data: Distributed and Parallel Architectures; Cloud Computing for Big Data.
Subject 3.
Storing and pre-processing where the following topics are addressed: Big Data Storage: Structured Storage; Non-Relational Database; Nosql Database Types; Big Data Preprocessing Techniques: Types of errors; Error handling. Tutorial on DB nosql
Subject 4.
Pre-processing techniques: where the following topics are addressed: Big Data pre-processing techniques - Filtering, Transformation, Integration. Practice on pre-processing techniques.
Subject 5.
Dimensional reduction: where the following topics are addressed: Size reduction of Big Data: Principal-Component Analysis, Singular-Value Decomposition. Tutorial on PCA and SVD.
Subject 6.
Big Data Clustering: Cluster Partitioning; Fuzzy Clustering; Relational Clustering. Practice on the clustering.
Subject 7.
Classification: where the following topics are addressed: Big Data Classification Algorithms: Classification Criteria; Bayesian Classifiers; Support Vector Machines.
Subject 6.
Prediction; where the following topics are addressed: Big Data Based Prediction Algorithms: Finite State Machines; Probability Models; Recurring Models.

Evaluation system and criteria

The exam consists of a written test to assess the ability to analyze and rework the acquired concepts and a series of activities (e-tivity) carried out during the course in virtual classes.
The expected learning outcomes about the subject's knowledge and ability to apply them are assessed by the written test, while the communication skills, the ability to draw conclusions and the ability to self-learn are evaluated in itinere through e-tivities.

Bibliography and resources

4. *Materials to consult:*

Notes written by the instructor are available in English. The notes cover the course contents and examination programme.

5. Recommended bibliography:

J. Leskovec, A. Rajaraman, J. D. Ullman, Mining of Massive Datasets