Dipartimento di Sociologia e Ricerca Sociale
Anno accademico 2022/2023

Big Data Technologies [145677]

No class division

Corso di studio Data Science
Ordinamento Data Science
Percorso standard

Docenti: DANIELE MIORANDI (Tit.), STEFANO TAVONATTI

Numero ore: 48

Periodo: Second semester

Crediti: 6

Settori: ING-INF/05

Course objectives and learning outcomes

Nowadays we are producing data at rates that we have never seen before, creating datasets characterized by extreme Volume, Variety and Velocity. Unfortunately, traditional data management technologies have been proven limited in managing data with these characteristics. This led to the term Big Data, as a way to refer to this kind of data, and new technologies been developed to cope with it. This course is an introduction to Big Data Technologies. It aims at providing an understanding of the fundamental principles, frameworks and tools upon which Big Data systems are built.
The students will acquire an understanding of what is big data, why it is relevant, knowledge of some of the technologies/frameworks/tools available for building big data systems and how to combine them to analyse data at scale

Entrance requirements

Basics of data and databases
Basics of programming
Working usage of Command Line Interface (CLI)

Contents

<ul><li>The big picture: tech megatrends </li><li>Data modelling: Data vs data representation; Structured vs unstructured data; Relational data model; Semi-structured data models; Examples: csv, json, xml etc.; Graph data models; Data model vs data format; Data streams; Batch vs stream processing (intro) </li><li>Characteristics of big data: The 3 (5) Vs, Big data vs Small data Getting value out of big data, Big data strategy </li><li>Security and privacy in big data Legal aspects of data: GDPR; Data licensing; Privacy in big data </li><li>Big data management systems: Relational DBs; No-SQL DBs </li><li>Storing big data: HDFS; Data warehouse; Data lake; Object storage </li><li>Big data retrieval: Querying SQL; Querying JSON; SPARQL </li><li>Big data ingestion: Ingestion infrastructure; Message queues; Pub/Sub; MQTT; Apache Kafka </li><li>Batch processing: MapReduce; Apache Spark </li><li>Stream processing: Spark Streaming; Apache Flink </li></ul>

Teaching and learning methods and activities

<p>Flipped classroom: material is shared before the class for students to prepare. During the lectures the contents are revised, discussed and applied. </p>

Tests and assessment criteria

Two options:

Option 1: project (mixed pairs A/B) [25/30] and written test [6/30]. Project results include a 5 pages report + working code.
Option 2: written test [6/30] and oral exam [25/30] - the latter includes a pen-and-paper project to be developed

Bibliography /study materials

Not applicable.

Other information

<p>The material used in the classes is published in: https://bit.ly/bdt-unitn-2023 </p><p>Notifications and notices are published on: https://t.me/joinchat/AAAAAFbf-MeERnNFDVRphg</p>

Stampa del 19/02/2023