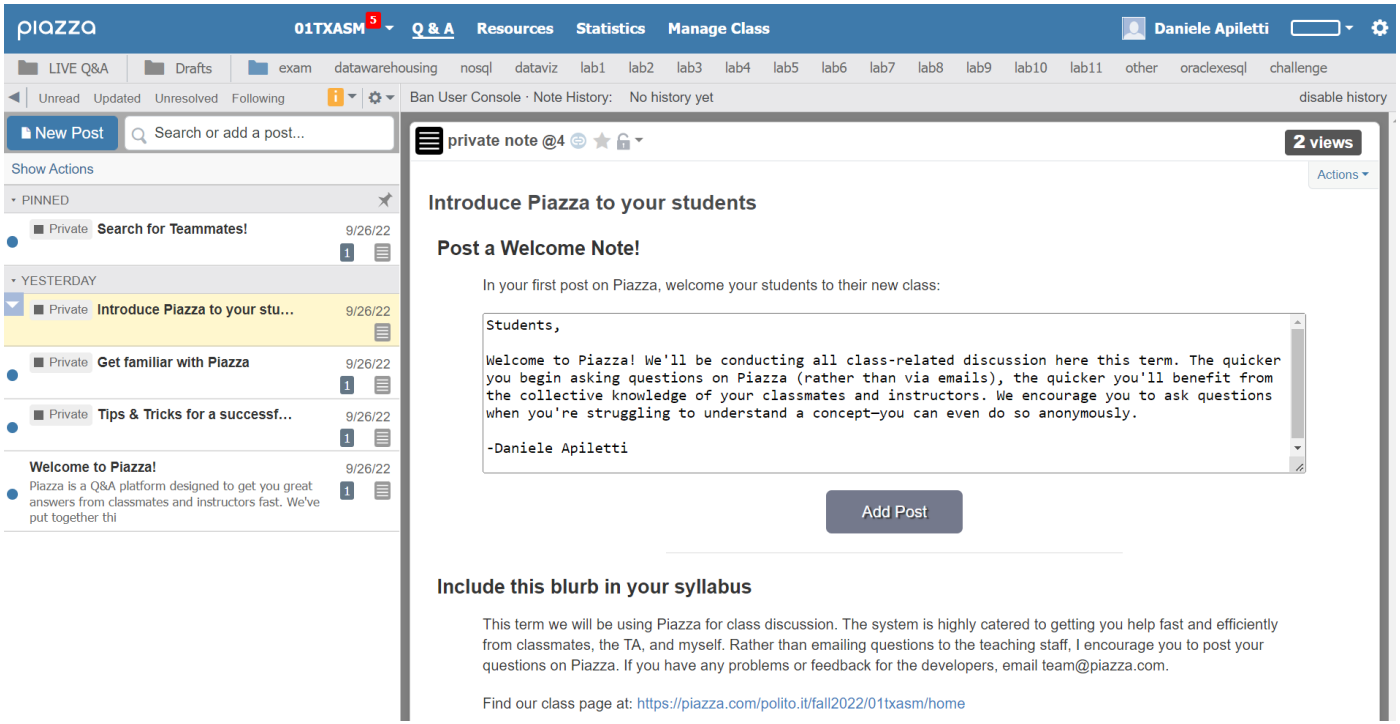
Big data: architectures and data analytics

Teachers

- Daniele Apiletti 
  - Main lecturer
- Simone Monaco & Luca Colomba 
  - Exercises
  - Laboratory practices
  - Student assistance

reach us by email: name.surname@polito.it
or better get assistance on Piazza:
https://piazza.com/polito.it/fall2023/01qydov

Piazza Q&A



We are using Piazza for class discussion. The system is highly catered to getting you help fast and efficiently from both classmates and teachers. Rather than emailing questions to the teaching staff, please post your questions on Piazza, even anonymously.

We might use Piazza for **announcements** in case of **failure** of either the Polito teaching portal or the Virtual Classroom services (or both).

Weekly schedule

	lunedì 16/10/2023	martedì 17/10/2023	mercoledì 18/10/2023	giovedì 19/10/2023	venerdì 20/10/2023
9 00					
10 00					
11 00			**Big data: architectures and...** APILETTI DANIELE AA - ZZ 3P Lezione/Esercitazione classroom **12**		
12 00					
13 00					
14 00					
15 00					
16 00				**Big data: architectures and...** APILETTI DANIELE AA - ZZ R1 Lezione/Esercitazione	**Big data: architectures and...** APILETTI DANIELE AA - ZZ
17 00					**Big data: architectures and...** APILETTI DANIELE AA - ZZ
18 00					
19 00					

Weekly schedule

- Lectures (45 hours)
  - Wednesday 11:30-14:30
  - Thursday 16:00-19:00
- Practices (15 hours)
  - Friday 16:00-17:30 Team 1 (A-L)
  - Friday 17:30-19:00 Team 2 (M-Z)
  - No lab activities during the first weeks (*)
    - The first Lab is on Friday, **October 20 (*)**

Practices

- We will provide you a specific account on the BigData@Polito cluster
  - https://jupyter.polito.it
  - https://hue.polito.it
- Detailed information will be provided next week
  - You will receive an email from the administrator of the cluster with username and password

Topics

- Lectures
  - Introduction to Big data
  - Hadoop
    - Architecture
    - **MapReduce programming paradigm**
  - Spark
    - Architecture
    - **Spark programs based on RDDs (Resilient Distributed Data sets) and Spark SQL (DataFrames and Datasets)**

Topics

- Data mining and Machine learning libraries for Big Data
  - **MLlib** (Apache Spark's scalable machine learning library)
- Streaming data analysis
  - **Spark Streaming**
- SQL databases for relational big data and NoSQL databases
  - Data models, Design, Querying

Topics

- Laboratory activities
  - Application development on Hadoop and Spark

Prerequisites / prior knowledge

- Object-oriented programming skills
  - **Java language (mandatory)**
- and basic knowledge of traditional database concepts (recommended)
  - Relational data model
  - SQL language

Material

- Web page
  - https://dbdmg.polito.it/dbdmg_web/index.php/2023/09/27/big-data-architectures-and-data-analytics-2023-2024/
  - Slides, exercises, lab activities, past exams, etc.
- Online lecture recordings (virtual classrooms)
  - on the Teaching portal

    https://didattica.polito.it

Books and Readings

- Reference books:
  - Matei Zaharia, Bill Chambers. Spark: The Definitive Guide (Big Data Processing Made Simple). O'Reilly Media, 2018.
  - Advanced Analytics and Real-Time Data Processing in Apache Spark. Packt Publishing, 2018.
  - Matei Zaharia, Holden Karau, Andy Konwinski, Patrick Wendell. Learning Spark (Lightning-Fast Big Data Analytics). O'Reilly, 2015.
  - Tom White. Hadoop, The Definitive Guide. (Third edition). O'Reilly Media, 2015.
  - Donald Miner, Adam Shook . "MapReduce Design Patterns: Building Effective Algorithms and Analytics for Hadoop and Other Systems." O'Reilly, 2012

Exam rules

- Written exam
  - 2 programming exercises (max 27 points)
    - Design and develop Java programs based on the Hadoop MapReduce programming paradigm and/or Spark RDDs
  - 2 questions / theoretical exercises (max 4 points)
    - Topics
      - Technological characteristics and architecture of Hadoop and Spark
      - HDFS
      - MapReduce programming paradigm
      - Spark RDDs, transformations and actions
      - Spark SQL
      - Spark Streaming
      - Spark MLlib
      - NoSQL databases and data models for big data

Exam rules

- On-site written exam on the Exam platform (Moodle) with Lockdown browser
  – **you must bring your own PC** –
  - 90 minutes
  - The exam is **open book**
    - Books, notes, and paper material are allowed
    - Electronic devices of any kind (PC, mobile phone, calculators, etc.) are not allowed, besides the PC used for the Exam itself.
- Past exams will be available to practice
- Students can fail…