CORSO: Architectures and systems for big data

DOCENTI: Paola Vocca (Ph.D., 1993);
EMAIL: paola.vocca@uniroma2.it
PAGINE WEB: http://dii.uniroma2.it/membro_personale/vocca-paola/
https://www.paolavocca.it/

DESCRIZIONE DEL CORSO

Il corso ha lo scopo di far acquisire allo studente una buona conoscenza dei principi che governano la progettazione di sistemi per il trattamento di ingenti moli di dati e permettono di modellarne il comportamento. Il Corso propone i concetti basilari delle architetture distribuite al servizio del processamento ed immagazzinamento di Big Data e li declina nello studio delle tecniche di preelaborazione, riduzione dimensionale, clustering, classificazione e predizione. Inoltre, obiettivo formativo del Corso è fornire allo studente una conoscenza nel dettaglio sugli approcci alla memorizzazione e strutturazione dei dati, sia relazionali che non-relazionali. Le esercitazioni associate al Corso sviluppano le competenze necessarie a progettare ed analizzare sistemi per i Big Data attraverso l'uso del calcolatore.

OBIETTIVI DI APPRENDIMENTO

Il Corso ha i seguenti obiettivi formativi:

- ✓ Illustrare i principi basilari dei sistemi di gestione di grandi moli di dati
- ✓ Illustrare i diversi approcci architetturali allo stato dell'arte
- ✓ Illustrare le principali tecnologie per la manipolazione dei dati

Ore di lezione frontale

METODOLOGIA

Il corso prevede ore ti lezioni teoriche ed ore di esercitazione in cui si farà uso di Apache Spark e Spark SQL.

VALUTAZIONE

Progetto individuale: 70%;
Orale 30%

PROGRAMMA

1. Introduzione ai concetti di base dei Big Data:
   1.1. terminologia, aspetti principali ed esempi di applicazioni.
   1.2. Problematiche principali legate alla gestione dei Big Data: volume dei dati e occupazione di memoria, velocità dell'elaborazione e complessità computazionale, presenza di errori e accuratezza dei dati, comprensione dei dati.
2. Architetture per la gestione dei Big Data:
   2.1. Architetture Distribuite e Parallele;
   2.2. Cloud Computing per i Big Data.
   2.3. **Esercitazione su Cloud per il Big Data**.
3. Paradigmi di programmazione per i big Data:
   3.1. Memorizzazione dei Big Data:
   3.2. Memorizzazione strutturata;

3.3. Database non relazionali; Tipologie di database NoSql; Tecniche di preelaborazione dei Big Data: Tipi di errori; Gestione degli errori. Esercitazione su DB NoSql.
3.4. Paradigma Hadoop e Map Reduce
3.5. Spark e SparkSQL
3.6. Spark SQL, Datasets and DataFrames
3.7. Spark MLlib
3.8. Spark Streamin

Esercitazione in Hadoop e Map Reduce e Spark SQL, Datasets and DataFrames

LIBRI DI TESTO

- ✓ Materiale fornito dal Docente
- ✓ **Big Data: Principles and Paradigms "** di Rajkumar Buyya, S. Thamarai Selvi, e Xingchen Chu (2016) - Questo libro offre una panoramica completa dei principi, dei modelli e delle tecnologie fondamentali relativi al campo del "Big Data".
- ✓ **Spark: The Definitive Guide** di Bill Chambers e Matei Zaharia (2018) - Questo libro si concentra su Apache Spark, un framework di elaborazione dei dati in memoria che è particolarmente efficace per l'elaborazione di grandi dataset