UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

SYLLABUS DEL CORSO

Technological Infrastructures for Data Science

2324-2-FDS01Q016

Obiettivi

Il corso ha lo scopo di fornire una solida conoscenza delle piattaforme tecnologiche (Cloud - Containerizzazione) che consentono la raccolta e la gestione di dati, nonché delle piattaforme elaborative (architetture, algoritmi e infrastrutture) utilizzate per analizzarli.

Le esercitazioni previste forniranno allo studente le competenze di base necessarie per interagire con tali piattaforme.

Contenuti sintetici

Piattaforme virtualizzate per la raccolta e gestione di dati di grandi dimensioni e velocità . Architetture per l'elaborazione dei dati, infrastrutture elaborative, piattaforme Big Data per la Data Science, esempi di piattaforme. Modelli e pratiche per lo sviluppo del software e modelli di ML: Waterfall, DevOps, DataOps, MLOps.

Programma esteso

1. Introduzione al corso

- Importanza del ruolo del Data Engineer nell'ambiente professionale
- Pipeline di dati
- Architettura di riferimento dei Big Data del NIST

2. Cloud Computing:

- La virtualizzazione come principale fattore abilitante del Cloud Computing
- Introduzione al cloud computing, compresi i modelli di servizio, i modelli di distribuzione e le caratteristiche.
- Approfondimento delle offerte cloud attuali, confrontando AMAZON EC3 e Azure web app.
- Informatica senza server

3. Containerizzazione:

- Introduzione ai container Linux e a Docker
- Esecuzione di un container
- Volumi: dati persistenti per i container
- Rete per i container
- Applicazioni multi-contenitore
- Gestione delle risorse multi-host per carichi di lavoro containerizzati con Kubernetes

4. Organizzazione e distribuzione dei dati:

- Concetto di lago di dati
- Magazzino di dati di grandi dimensioni (DW)
- HDFS (Hadoop Distributed File System), Avro, Parquet (formati di archiviazione dei dati).

5. Piattaforme di elaborazione dei Big Data:

- Gestione delle risorse con Apache YARN
- Basi dell'elaborazione batch e Apache Spark
- Nozioni di base sull'elaborazione dei flussi, Apache Storm e Spark Streaming.

6. Processo di sviluppo del software:

- Servizi e Service Computing
- Introduzione all'ingegneria del software
- Metodologie di sviluppo Waterfall e Agile
- Concetti di DevOps e DataOps

Prerequisiti

Conoscenza di base di:

- un linguaggio di programmazione (es. Python)
- dell'architettura di un calcolatore.

Modalità didattica

Lezioni ed esercitazioni in aula. Il corso verrà erogato in lingua inglese

Materiale didattico

Dispense e slide del corso fornite dai docenti.

Periodo di erogazione dell'insegnamento

Secondo anno, primo semestre

Modalità di verifica del profitto e valutazione

L'esame sarà costituito da due parti. Le due parti che dovranno essere **svolte nel medesimo appello**.

La prima parte consisterà in un insieme di domande chiuse ed aperte da svolgersi in un'ora (orientativamente si avranno 9 domande chiuse ed 4 aperte, tuttavia si potranno avere piccole variazioni nella struttura dell'esame). La prima parte dell'esame si svolgerà in modalità elettronica (esami online) ed il peso di ciascuna domanda per la formazione del voto verrà esplicitamente indicato contenente le domande. L'esame verrà erogato in lingua inglese, lo studente ha la facoltà di rispondere in inglese o in italiano.

La seconda parte sarà costituta da un approfondimento di un tema concordato con il professore da realizzare in gruppi di 2 persone e dalla sua esposizione orale.

Esempi di tipi di progetto:

- Analisi e test di una particolare piattaforma tecnologica (vantaggi, svantaggi, costi, curva di apprendimento, realizzazione di un'applicazione che utilizzi tale tecnologia)
- Progettazione di un'applicazione per il cloud (scelta del provider, del tipo di macchine virtuali, dei servizi, stima dei costi, stima della qualità del servizio, stima delle dimensioni dei dati supportati, attività correttive necessarie in caso di imprevisti)

Una volta che lo studente avrà svolto entrambe le prove, l'esame si considererà superato se si verificheranno entrambe queste condizioni:

1. Per entrambe le parti lo studente avrà ottenuto più della metà dei 15 punti a disposizione.
2. La somma dei punti delle due parti sarà maggiore o uguale a 18.

In tal caso lo studente potrà registrare un voto costituito dalla somme dei punti.

Orario di ricevimento

Martedì 12:30-14:30, chiedere conferma per email

Sustainable Development Goals

IMPRESE, INNOVAZIONE E INFRASTRUTTURE