UNIVERSITÀ
DEGLI STUDI DI MILANO-BICOCCA

SYLLABUS DEL CORSO

Technological Infrastructures for Data Science

2425-2-FDS01Q016

Obiettivi

Il corso ha come scopo principale quello di fornire allo studente una solida e fondativa conoscenza delle principali **soluzioni tecnologiche** e **metodologie di sviluppo software** a supporto della data science. Le esercitazioni forniranno allo studente le competenze di base necessarie per interagire con tali strumenti.

Contenuti sintetici

Il corso è costituito dai seguenti moduli:

Modulo 1 - Infrastruttura: Introduzione alla virtualizzazione, Cloud Computing e Containerization

Modulo 2 - Piattaforma: Organizzazione e distribuzione dei dati, Data Lake, HDFS, YARN

Modulo 3 - Processamento: Batch vs Streaming vs Messaging, i casi di Hadoop, Spark, Storm, Kafka

Modulo 4 - Sviluppo software: Waterfall, Agile, DevOps, DataOps, MLOps

Programma esteso

Argomenti del corso divisi per moduli:

Modulo 1 - Infrastruttura:

- La figura del data engineer

- L'architettura di riferimento
- Virtualizzazione
- Cloud Computing (Introduzione, Modelli di servizio e di deployment, cararatteristiche essenziali)
- Containerization con Docker
- Serverless

Modulo 2 - Piattaforma:

- Il Data Lake
- HDFS e YARN

Modulo 3 - Processamento:

- Batch processing (Apache Hadoop e Apache Spark)
- Stream processing (Apache Storm, Apache Spark e Apache Flink)
- Messaging (Apache Kafka)

Modulo 4 - Sviluppo Software:

- Service computing
- Software engineering
- Metodologie di sviluppo (Waterfall, Agile, DevOps, DataOps, MLOps)

Prerequisiti

Conoscenza basilare dell'architettura di un computer (CPU, RAM, Storage), Sistemi operativi, della Shell dei comandi, del linguaggio di programmazione Python e Jupyter.

Modalità didattica

Insegnamento con differenti modalità didattiche:

- 15 lezioni da 2 e 3 ore svolte in modalità interattiva in presenza; In tale modalità comunque il docente coinvolgerà gli studenti per mezzo di domande e spunti.
- 6 laboratori da 3 ore svolti in presenza - modalità interattiva.

Il corso verrà erogato in lingua inglese

Materiale didattico

Dispense e slide del corso fornite dai docenti.

Si segnalano i seguenti testi per approfondimento:

- The basics of cloud computing ISBN-13: 978-0124059320 Autori: Derrick Rountree, Ileana Castrillo
- Designing Data-Intensive Applications: The Big Ideas Behind Reliable, Scalable, and Maintainable Systems

ISBN-13: 978-1449373320 Autore: Martin Kleppmann

Periodo di erogazione dell'insegnamento

Secondo anno, primo semestre

Modalità di verifica del profitto e valutazione

La valutazione è costituita da due parti: una **prova scritta** e una **discussione orale** su un argomento di approfondimento.

La prova scritta consiste in domande aperte e chiuse sugli argomenti del corso. Questa ha una durata di circa un'ora ed un punteggio massimo di **17 punti**.
Una tipica prova scritta comprende 12-13 domande a scelta multipla e 2-3 domande a risposta aperta, per le quali si hanno a disposizione dai 60 ai 75 minuti.
La prova orale consiste nella discussione di una tematica non trattata durante il corso o di un approfondimento di una tematica trattata. Il lavoro di approfondimento (lavoro ricerca e creazione delle slide) potrà essere realizzato in gruppi di massimo 3 persone ma la discussione e la valutazione sono personali.
L'argomento dell'approfondimento deve essere previamente accordato con il professore. La prova orale dà diritto ad un massimo di **15 punti**.

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

- per entrambe le parti lo studente avrà ottenuto più di (o uguale a) 7 punti
- la somma dei punti delle due parti sarà maggiore o uguale a 18

In tal caso potrà essere registrato un voto costituito dalla somme dei punti.

Non è prevista l'erogazione di prove parziali durante il corso.

Orario di ricevimento

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

Sustainable Development Goals

IMPRESE, INNOVAZIONE E INFRASTRUTTURE