UNIVERSITÀ DI PISA

DISTRIBUTED DATA ANALYSIS AND MINING

ROBERTO TRASARTI

Anno accademico	2023/24
CdS	DATA SCIENCE AND BUSINESS INFORMATICS
Codice	687AA
CFU	6

Moduli	Settore/i	Tipo	Ore	Docente/i
DISTRIBUTED DATA ANALYSIS AND MINING	INF/01	LEZIONI	48	ROBERTO TRASARTI

Prerequisiti (conoscenze iniziali)

- Data Mining I e II
- Programmazione Python di base

Programma (contenuti dell'insegnamento)

Il Data Mining sui Big data è oggi un'area di ricerca molto attiva. L'applicazione delle attuali metodologie analitiche e strumenti software su un singolo personal computer non può gestire in modo efficiente dataset di grandi dimensioni. Le piattaforme di calcolo distribuito sono una soluzione scalabile per il big data mining, attraverso la scomposizione del problema in operazioni più piccole che possono essere eseguite parallelamente su singoli processori / macchine. Il corso propone l'insegnamento di concetti base del paradigma di calcolo distribuito tramite MapReduce dal punto di vista teorico e pratico, in particolare ci si focalizzerà su Hadoop per lo sviluppo di competenze nell'uso di strumenti di calcolo ad alte prestazioni per il data engineering, l'analisi di dati e l'utilizzo di tecniche di data mining. Gli studenti impareranno come i classici algoritmi di data mining possono essere applicati sui Big Data usando Hadoop (Spark). Set di dati reali (e open source) verranno utilizzati per presentare esempi e per consentire agli studenti di costruire i propri progetti. Una metà delle lezioni consisterà in esercitazioni (laboratorio) e una metà delle lezioni sarà teorica.

- Motivations: What is and Why Distributed Data Mining is needed in a Big Data Scenario
- Recall parallel and distributed computing notions
- Amdahl's law, differences between shared and distributed memory architectures
- Introduction to Hadoop
- Hadoop Ecosystem
- Interacting with HDFS
- Hadoop Combiners
- Basic Spark and RDD
- Map-Reduce Programming Patterns
- Recall Python programming
- Data Analysis with Spark
- Data Mining and Machine Learning with Spark
- Spark SQL
- Spark Streaming
- Example on how to prepare a project

Indicazioni per non frequentanti

https://teams.microsoft.com/l/channel/19%3anFtfCCGn_6R4UsBgqJ9_jnpd9V-diJsGsiHo_3Gb9481%40thread.tacv2/Generale?groupId=d498e90a-6ea1-41ba-9f0f-682b33ffdc95&tenantId=c7456b31-a220-47f5-be52-473828670aa1

Modalità d'esame

Students groups composed by 3-5 students to develop a project (report + short slide presentation);

UNIVERSITÀ DI PISA

Note

INIZIO LEZIONI 18/09

Materiale del Corso: https://teams.microsoft.com/l/channel/19%3anFtfCCGn_6R4UsBgqJ9_jnpd9V-diJsGsiHo_3Gb9481%40thread.tacv2/Generale?groupId=d498e90a-6ea1-41ba-9f0f-682b33ffdc95&tenantId=c7456b31-a220-47f5-be52-473828670aa1

Ultimo aggiornamento 17/09/2023 16:08