Course Title	Big Data Analytics				
Course Code	AI615				
Course Type	Compulsory				
Level	Master (2nd cycle)				
Year / Semester	1st Year / 2nd Semester				
Teacher's Name	TBA				
ECTS	8	Lectures / week	Up to 6 Teleconferences	Laboratories / week	None
Course Purpose and Objectives	Big Data requires the storage, organization, and processing of data at a scale and efficiency -typically of heterogeneous nature and in streaming flow- that go well beyond the capabilities of conventional information technologies. Such requirements have been first introduced for processing the web, and they are today a common place in many industries. In modern enterprise systems, data lives in many different places (e.g., in relational databases, in `data lakes' on distributed file systems, behind REST APIs, or is constantly being scraped from web resources), and comes in many different formats. This reality calls for novel query and programming interfaces (e.g., Map/Reduce) and computing models (As_A_Service) in enterprise premises as well as on the Cloud. This course aims in a first step to introduce parallel/distributed data processing using the MapReduce (M/R) paradigm and provide insights for developing applications on top of popular Big Data platforms such as Spark. Big data raises also new challenges in data mining. Given the scale and speed of data that needs to be processed as well the variety of parameters to be taken into account, state of the art machine learning algorithms working offline and expecting homogeneous and clean data are also challenged. There is on ongoing effort to design Big Data Mining algorithms accommodating a parallel/distributed or even a streaming evaluation. Of course such kind of incremental, partial evaluation impacts the quality of obtained statistical models and thus algorithms compromise between quality of the learning and computation time. This course aims to present Big Data Mining techniques used in real applications by adopting an algorithmic viewpoint: data mining is about applying algorithms to data, rather than using data to "train" a machine-learning engine of some sort.				
Learning Outcomes	Upon succesful completion of this course students should be able to: • *Understand* different models of computation: a. MapReduce b. Streams and online algorithms • *Mine* different types of data: a. Data is high dimensional				

	b. Data is infinite/never-ending • *Use* different mathematical 'tools': a. Hashing (LSH, Bloom filters) b. Dynamic programming (frequent itemsets) • *Solve* real-world problems: a. Duplicate document detection b. Market Basket Analysis		
Prerequisites	None	Co-requisites	None
Course Content	1: Introduction on Big Data Processing & Analytics. The challenges of Big Data Mining 2-3: The MapReduce Programming Paradigm The MapReduce Software Ecosystem (Hadoop and SPARK) 4: Data Lakes and Warehouses Relational Query Processing Algorithms on MapReduce 5: NoSQL Database Systems MongoDB, Hive, etc. 6: Finding Similar Items Minhashing and Locality-sensitive Hashing 7: Extracting Association Rules Centralized and Distributed Mining of Frequent Patterns 8-9: Analysing Data Streams Sampling, Windows, Synopses and Sketches 10-11: IoT Data Anaytics Anomaly Detection for Batch and Stream Processing 12: Big Data Entity Resolution Indexing and Matching Semistructured Heterogeneous Entities 13: Big Data Analytics Accountability Fairness, Diversity, Transparency and Neutrality		
Teaching Methodology	E-Learning		
Bibliography	Jure Leskovec, Anand Rajaraman, Jeff Ullman. "Mining of Massive Datasets" Cambridge University Press, Latest Edition Entity Resolution in the Web of Data (Synthesis Lectures on the Semantic Web: Theory and Technology), Morgan & Claypool Publishers, Latest Edition Uri Laserson. Advanced Analytics With Spark: "Patterns for Learning from Data at Scale" O'Reilly Media , Latest Edition Gerard Maas, Francois Garillot. "Learning Spark Streaming" O'Reilly Media, Latest Edition Donald Miner, Adam Shook "MapReduce Design Patterns" O'Reilly Media, Latest Edition		

Assessment	Final Examination	50%
	Assignments/On-going evaluation	50%
		100%
Language	English	