BigData: Big Data, Data Intensive Technologies and Analytic Techniques

Syllabus and Procedures

Description:

Big Data, also referred to as Data Intensive Technologies, are becoming a new technology trend in science, industry and business. Big Data are becoming related to almost all aspects of human activity from just recording events to research, design, production and digital services or products delivery to the final consumer. Current technologies such as Cloud Computing and ubiquitous network connectivity provide a platform for automation of all processes in data collection, storing, and processing. It is beyond Human capabilities to just inspect this data and make sense of it. Therefore, we must use automated analytical techniques to extract useful information from this data, which we can use visualise, understand or perform automated predictions using this data.

Aims:

The aim of this module is to cover both the conceptual and architectural issues in defining the Big Data infrastructure and data analytics applications and more specific information about Big Data ecosystem components, tools, and analytic techniques that can be applied to this data. The module provides detailed analysis of the Big Data use cases in science, industry and business what is used to motivated better understanding of the Big Data technologies and required solutions. The module uses two major components around the whole curriculum is built: conceptual Big Data definition and Architecture Framework for Big Data Ecosystem that includes five components: data models and structures; data management; analytic techniques and tools, cloud based Big Data Infrastructure, Big Data Security, data source/origin and data target/use. The module provides extended security analysis of the Big Data security services and privacy issues. In addition, the module provides an introduction to Analytics/Data Mining/Machine Learning, and covers both the theoretical background and practical application of these techniques. In the last seminar, the module proposes an approach to enterprise data processes analysis and their possible optimisation using Big Data technologies.

Syllabus

This module provides students with a clear picture of the emerging Big Data and Data Intensive technologies and Data Analytics techniques that power the practical use of Big Data. Big Data consolidate recent achievements in Cloud Computing, High Performance Computing, Web/Internet technologies, mobile and sensor networks. It gives students a chance to experience a variety of Big Data technologies and related analytical techniques. By the end of this module, students will have studied:

1. Big Data and Data Intensive technologies definition; Big Data properties: Volume, Velocity, Variety, Variability, Value, Veracity.
2. Major use cases in science, industry, Internet or web, and their Big Data challenges.
3. Big Data Architecture Framework: main components and their inter-relation
4. Big Data Infrastructure and its cloud based implementation, including Federated Access and Delivery Infrastructure.
5. Data structures, file systems for distributed data storage and processing; new database architectures for big data (noSQL, key-value storage, in-memory, etc).
6. Big Data Analytics tools and platforms: batch, real-time/streaming, near real-time; Open Source Analytics platforms and from major cloud and business analytics providers.
7. Major Big Data analytics methods: statistical analysis, cluster analysis, graph analytics, Bayesian analysis, machine learning; as well as market oriented analytics.

8. Simple statistical techniques that can be used for the analysis of data.
9. Extracting knowledge from datasets to further the understanding of business processed, using techniques such as rule induction and cluster analysis.
10. Using Machine Learning techniques for making predictions, such as Neural Networks and Support Vector Machines.
11. Big Data security and data protection; data centric security models; Big Data privacy issues, differential privacy and data re-identification.
12. Data protection and corresponding standards, regulation and legislation; data provenance.
13. Enterprise business processes and data management issues; analysis, evaluation and optimisation; enterprise data applications engineering.

Text Book:

The following two textbooks are considered for the 'analytics' part of the module.

Data Science for Business (2013): What you need to know about data mining and data-analytic thinking. Author Foster Provost. Publisher: O'Reilly Media; 1 edition (August 16, 2013) **ISBN-10:** 1449361323. **ISBN-13:** 978-1449361327
Data Mining for the Masses, available for download in pdf format from the RapidMiner website
Students can also download the free textbook

Other References:

1. Standards and BCP documents from NIST, CSA, others
2. Whitepapers, Tutorials and documentation from Amazon AWS, Microsoft Azure

Lists of additional books, articles and sites links will be provided during the course.
Videos: A range of training videos are available from the RapidMiner website.

Recommended Prior Knowledge:

It is recommended that the students have knowledge of computer architecture, operating systems and network technologies. Completion of modules on Computer Systems and Operating Systems are mandatory pre-requisites. Module on databases, data management or business process management are recommended.
For the analytics part of this module, no programming is required, as we will be using a free analytics workbench called RapidMiner. You will require basic mathematical skills (numeracy), and must be familiar with using a spreadsheet tool such as Excel.

Recommended Preparatory Reading:

The Fourth Paradigm: Data-Intensive Scientific Discovery. By Jim Gray, Microsoft, 2009. Edited by Tony Hey, et al. [online] http://research.microsoft.com/en-us/collaboration/fourthparadigm/

Syllabus:

Seminar 1: Introduction. Big Data technology domain definition

Topics: Big Data and Data Intensive technologies definition; Big Data properties: Volume, Velocity, Variety, Variability, Value, And Veracity. Big Data ecosystem, data origin, data target; raw data and actionable data.

Seminar 2: Big Data use cases from Science, industry and business

Topics: Big data use cases analysis from data intensive science: LHC (Large Hadron Collider), LOFAR/SKA (Square Kilometer Array) astronomy, genomic research. Identification of Big Data challenges and requirements. “Long tail” science and challenges.
Big data use cases analysis from industry, Internet or web: target advertisement, recommender system (e.g. Netflix), targeted campaigns, network security and intrusion detection. Identification of Big Data challenges and requirements.

Seminar 3: Architecture Framework for Big Data Ecosystem, Big Data Infrastructure components
Topics: Big Data Architecture Framework: main components and their inter-relation. General purpose Big Data Infrastructure and its cloud based implementation, including Federated Access and Delivery Infrastructure.
File systems for distributed data. Data storage types; new database architectures for Big Data (noSQL, key-value storage, in-memory storage/database, etc). Available open datasets and their use: Wikipedia, Twitter, AWS genome.

Seminar 4: Big Data analytic techniques, introduction to RapidMiner
Topics: Big Data Analytics tools and platforms: batch, real-time/streaming, near real-time. Open Source Analytics platforms and from major cloud and business analytics providers. Overview Big Data structure and models [TB: OK]. Introduction to analytics and different analytical techniques. What are Analytics/Data Mining/Machine Learning? Introduction to the RapidMiner toolkit. The use of simple pre-processing and Statistical techniques for modeling data will be described and implemented using RapidMiner.

Seminar 5: Understanding the processes behind Big Data.
Topics: Introduction to Rule Extraction Algorithms and Cluster Analysis. Evaluating data from text streams on the Internet. This seminar will describe techniques for the extraction of Knowledge from datasets, and problems that can be encountered when using these techniques. Decision tree induction and Cluster Analysis techniques will be described, and how to use these techniques to enhance understanding of business processes.

Seminar 6: Classification and forecasting techniques.
Topics: This seminar introduces a two popular Machine Learning technique, Neural Networks and Support Vector Machines. The seminar will discuss how they can be used for business processes such as classification and forecasting. In addition, it will describe several techniques used to analyze the performance of our analytic processes. It discusses how the performance technique selected must be matched with the business process being modeled. Measurement techniques such as Receiver Operating Curves and Gains Charts (sometimes called 'Lift' charts) can be used to select the best model from a range of models.

Seminar 7: Big Data Security, Protection, and Privacy
Topics: Big Data security and data protection; data centric security models. Big Data privacy issues, differential privacy and data re-identification.

Seminar 8: Integrating Big Data applications into enterprise IT infrastructure, data regulation compliance
Topics: Enterprise business processes and data management issues; analysis, evaluation and optimisation; enterprise data applications engineering. Data protection and corresponding standards, regulation and legislation; data provenance.