Module Name: (B.4) Algorithms and Systems for Big Data Processing

Aim

The efficient processing, storage and transmission of Big Data imposes significant challenges. The module aims to introduce some modern techniques, systems and platforms for effective Big Data analysis and processing. Since the topic of Big Data is rather broad, the module concentrates mainly on technologies developed within the context of the Apache Spark open source system. Spark is a modern and effective system for distributed Big Data Processing and within its context it is possible to study many problems related to Big Data and confront them effectively. At the context of Big Data, deep learning techniques are particularly effective and the module will introduce some of them. With the module, students will gain considerable experience on storage, processing and analysis of Big Data.

Learning Objectives

The main learning objectives include the ability to analyze, design and implement systems for Big Data storage, processing and analysis.

Learning Outcomes

On successful completion of this module, students should be able to:

- Understand the fundamentals of Big Data: Basic Concepts of Big Data (Volume, Variety, Velocity, Veracity, Validity and Volatility). Applications of Big Data with emphasis on Bioinformatics, usage case studies, open research problems, requirements for Big Data processing platforms.
- Familiarize with basic Big Data Processing Principles: Scaling, Efficiency, Fault tolerance, MapReduce/Hadoop, Hadoop Distributed File System (HDFS), Spark core, Spark SQL, Spark Machine Learning.
- Understand the principles of real time Big Data processing: real time stream processing, real time data processing, in memory data processing, Spark real time streaming with DStreams and with the new structured stream API.
- Exploit some effective approaches for Big Data: The Resilient Distributed Data Sets (RDDs) of Spark and their implementation, Application Development with RDDs, Spark high level API (DataFrames, DataSets, Spark SQL), Cluster Computing on Spark (standalone cluster management, Apache YARN, Apache Mesos, Cloud-based deployments).
- Introduce Big Data Analytics at the framework of Spark Machine Learning (Spark Mllib and Spark ML): Feature Extraction, Dimensionality Reduction, Principal Components Analysis, Binary and Multiclass Classification, Clustering techniques, Bayesian Inference, Text Analysis, Intoduction to deep learning and its applications for Big Data.

Bibliography

[1] Rajkumar Buyya, Rodroigo N. Calheiros, “Big Data: Principles and Paradigms”, Morgan Kafmann, 2016

[2] [2] Jules J Berman, “Principles and Practice of Big Data: Preparing, Sharing and Analyzing Complex Information”, 2nd edition, Academic Press, 2018

[3] [3] Md. Rezaul Karim, Sridhar Alla,, Scala and Spark for Big Data Analytics, Packt Publishing, 2017

[4] [4] Ian Foster, Dennis B. Gannon, William Grop, Ewing Lusk, Rich Wolski, Stig Telfer, “Cloud Computing for Science and Engineering (Scientific and Engineering Computation)”, MIT Press, 1st edition, 2017

[5] [5] Ian Goodfellow, Yoshua Bengio, Aaron Courville, “Deep Learning”, MIT Press, 2016

[6] [6] Kai Hwang, Min Chen, “Big-Data Analytics”, Wiley, 2017

[7] [7] Bill Chambers, Matel Zaharia, “Spark: The Definite Guide: Big Data Processing Made Simple”, O’Reilly, 2018