Big Data Technologies

Course Description

Data are often considered the “new oil”, the raw material from which value is created. To harness the power of data, the data need to be stored and processed on a technical level. This course introduces the four “Vs” of data, as well as typical data sources and types. This course then discusses how data are stored in databases. Particular focus is given to database structures and different types of databases, e.g., relational, noSQL, NewSQL, and time-series.Beyond classical and modern databases, this course covers a wide range of storage frameworks such as distributed filesystems, streaming, and query frameworks. This is complemented by a detailed discussion of data storage formats ranging from classical approaches such as CSV and HDF5 to more modern approaches like Apache Arrow and Parquet.Finally, this course gives an overview of distributed computing environments based on local clusters, cloud computing facilities, and container-based approaches.

Contents

1. Data Types and Data Sources
   1. The 4Vs of data: volume, velocity, variety, veracity
   2. Data sources
   3. Data types
2. Databases
   1. Database structures
   2. Introduction to SQL
   3. Relational databases
   4. nonSQL, NewSQL databases
   5. Timeseries DB
3. Modern data storage frameworks
   1. Distributed Filesystems
   2. Streaming frameworks
   3. Query frameworks
4. Data formats
   1. Traditional data exchange formats
   2. Apache Arrow
   3. Apache Parquet
5. Distributed Computing

5.1 Cluster-based approaches

1. Containers
2. Cloud-based approaches