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. The course discusses the most common data storage formats encountered in modern systems, focusing on both text­-based and binary data formats. Handling large amounts of data poses significant challenges for the underlying infrastructure. The course discusses the most important distributed and streaming data handling frameworks that are used in leading-edge applications.

Contents

1. Data Types and Data Sources
   1. The 4Vs of Data: Volume, Velocity, Variety, and Veracity
   2. Data Sources
   3. Data Types
2. Text-Based and Binary Data Formats
   1. Simple Formats: CSV, YAML
   2. XML
   3. JSON
   4. Hierarchical Data Format 5 (HDF 5)
   5. Apache Parquet
   6. Apache Arrow
3. NoSQL Data Stores
   1. Introduction and Motivation
   2. Approaches and Technical Concepts
4. Distributed Systems
   1. Hadoop and MapReduce
   2. Hadoop File System (HDFS)
   3. Spark
   4. DASK
5. Streaming Frameworks
   1. Spark Streaming
   2. Kafka