Using big data in practice: Computational infrastructures

Dr Gianluca Demartini

Big Data is challenging to process due to its volume and velocity dimensions. It is also difficult to make sense from it because of its variety and veracity challenges. In this talk we will discuss modern computational infrastructures to scale-out the processing of large datasets by means of distributed computing. We will introduce the concept of Map/Reduce and present Big Data systems like Apache Spark which is currently the market leader in Big Data processing solutions. We will explain by means of examples how large datasets can be processed in batch or as a stream.