Big Data at TU Wien: current deployment status and outlook

Linus Kohl[a], **Giovanna Roda**[a], **Siegfried Höfinger**[a,b], **Dieter Kvasnicka**[a,b]

[a] *TU Wien, TU.it*
[b] *VSC Research Center*

Hadoop, one of the major platforms for storage and processing of large scale datasets, provides a high performance, fault tolerant, scalable solution that can run on clusters of commodity hardware. What is also known as the Hadoop ecosystem (Fig. 1) is a continuously growing collection of software components, each of which covers different use cases and specialised tasks.

The TU Wien operates since December 2017 a little Big Data cluster that is being used for teaching and research. The cluster runs an open source Cloudera Hadoop distribution on 20 nodes with 256Gb of memory, 500TB HDFS and 300TB NFS storage and relies on a 10GB Ethernet interconnection. The Cloudera distribution offers an integrated environment for deploying Hadoop ecosystem components as well as an own cluster management service. Additionally, a smaller cluster with eight nodes is in place for testing purposes.

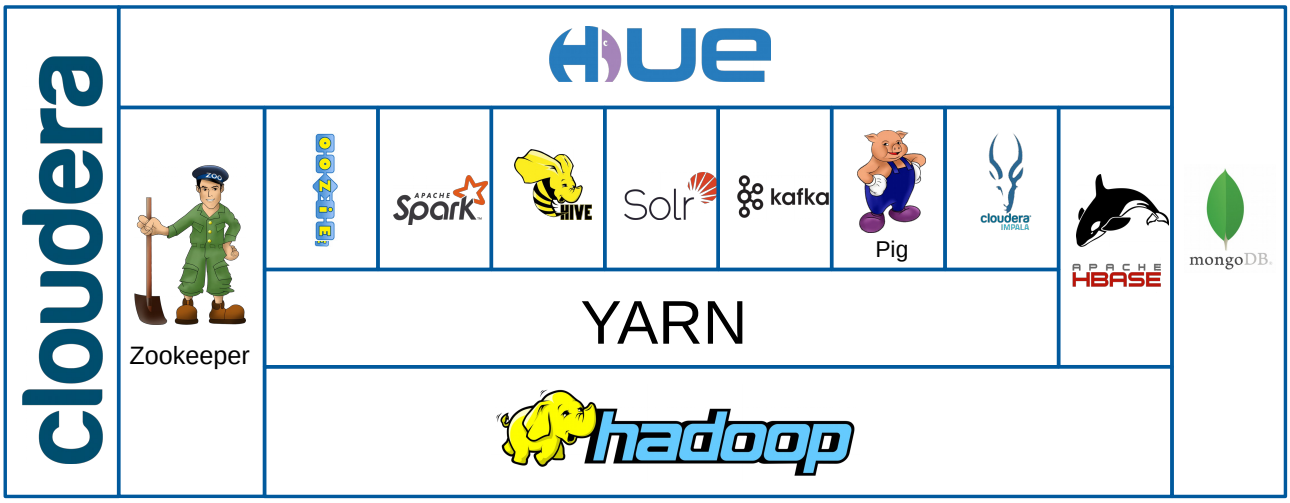

Fig. 1: The Hadoop ecosystem at TU Wien

The Data Lab Team manages the Hadoop cluster as part of the IT Services of the TU Wien. In this talk we will talk shortly about Hadoop and HDFS, its ecosystem, and will give a short outlook on Hadoop 3 and erasure coding, a data resiliency technique that reduces storage overhead from the 200% of the classical 3x replication, to only about 50%.

Furthermore, following the TU.it mission of increasingly addressing emerging R&D initiatives, the Data Lab is in the process of purchasing a new GPU cluster for Deep Learning. This extension of the Hadoop Cluster is going to support Deep Learning activities within the Faculty of Informatics. The initial installation will provide a prototype configuration of several state-of-the-art GPU servers forming the basis for DL workflows while retaining all the flexibility with respect to anticipated upgrades/reconfiguration in the immediate future.