Tárgytematika / Course Description

Big Data

GKNM_MSTA046

Tárgyfelelős neve / Teacher's name: dr. Takács Gábor
Félév / Semester: 2022/23/2
Beszámolási forma / Assesment: Vizsga
Tárgy heti óraszáma / Teaching hours(week): 2/2/0
Tárgy féléves óraszáma / Teaching hours(sem.): 0/0/0

OKTATÁS CÉLJA / AIM OF THE COURSE

The goal of the course is to teach modern approaches that are able to handle huge volumes of data. The curriculum covers the topic of distributed storage formats, task graphs (using the Dask library), stream processing (using the Streamz library), large scale machine learning and NoSQL databases.

TANTÁRGY TARTALMA / DESCRIPTION

- What is Big Data? Misconceptions. Challenges.
- The Dask library. Task graphs. Schedulers. Distributed arrays. Distributed data frames.
- The Parquet storage format. The Ubiquant data set. Basic queries. Comparison of different different approaches.
- Distributed linear regression.
- The stream processing paradigm. The Streamz library. Flow control. Branching and joining.
- Streamed linear regression.
- NoSQL databases.
- Distributed gradient boosting.

SZÁMONKÉRÉSI ÉS ÉRTÉKELÉSI RENDSZERE / ASSESMENT'S METHOD

Each student is assigned a Big Data related topic in the 7th week. Then they prepare a presentation about the topic and show in the 13th week (resit is done in the 14th). The mark is given based on the content and the quality of the presentation.

KÖTELEZŐ IRODALOM / OBLIGATORY MATERIAL

- Martin Kleppmann: Designing Data-Intensive Applications, O'Reilly, 2015.
- Dask Documentation (https://docs.dask.org/)