Big data

Motivace pro realizaci předmětu

Vývoj internetových aplikací je charakterizován neustálým zvětšováním objemu zpracovávvaných dat. Internetové portály, obchody, banky, pojišťovny, nemocnice, mobilní operátoři a další ve svých úložištích akumulují každý den velká množství dat. Data obsahují mnoho informací o zákaznících, o provozu, o produktech a proto podniky hledají efektivní metody jak tyto informace extrahovat a analyzovat, jaký hardware a jaké algoritmy použít pro zpracování perzistentních dat, ale i dat která jsou průběžně vytvářena a stále se mění, např. data ze sociálních sítí. V neposlení řadě se rozvíjejí metody jak data snadno komunikovat a prezentovat, jak na základě těchto informací lépe rozhodovat a řídit podniky. Algoritmy pro analýzu, rozhodování a předpovídání na základě učení jsou tradiční disciplínou umělé inteligence. Big Data je další vývojový krok ve zpracování dat. Techniky zpracování dat jsou proto velmi důležité pro řadu korporací a budou v budoucnu nezbytné pro všechny společnosti a podniky. Tyto techniky jim umožní si zachovat kompetitivní pozici na trhu. Kdo nebude data automaticky analyzovat a na základě těchto analýz kompetentně rozhodovat nebude konkurenceschopný.

Cílem předmětu je seznámit studenty s těmito novými trendy a technologiemi, neboť v praxi se s nimi budou stále častěji setkávat. IBM jako jedna z nejvýznamnějších světových technologických firem přispěla grantem na přípravu nového předmětu s touto tématikou. Motivace je zcela jasná: podle IBM analýzy bude v příštích několika letech na trhu velký nedostatek odborníků se znalostmi pro zpracování a analýzu velkých dat, poptávka bude velmi přesahovat nabídku. Věříme, proto že včasným zavedením předmětu, který se bude touto tematikou zabývat, pomůžeme absolventům k snadnému uplatnění na trhu práce. Pro realizaci předmětu využijeme vlastních bohatých zkušeností s vývojem podobných aplikací, na kterých spolupracujeme s průmyslem.

Big Data bude navazovat na předměty Strojové učení a analýza dat a Vytěžování dat v tom smyslu, že ukáže jak metody strojového učení aplikovat na velké soubory dat.

Přínos předmětu do profilu absolventa

Předmět se věnuje primárně technologiím pro zpracování velkých dat na počítačových klastrech. Představuje současná řešení a algoritmy, které podporují zpracování jak streamovaných tak perzistentních data s cílem analýzy, predikce a vizualizace. Absolventům nabídne standardní postupy ve formě praktických cvičení.

Rozsah předmětu

- 1+1; předpoládáme střídání přednáška - cvičení po týdnech

- Počet předpokládaných kreditů za absolvování předmětu: 2
- Počet studentů je maximálně 40. Ke každé přednášce a příslušnému tematickému bloku bude následovat cvičení.

Požadavky na studenty

Předmět je doporučován v letním semestru magisterského studia. Od studentů se očekává základní znalost programování v jazyce Java, základní znalosti z databází a základy optimalizace v rozsahu poskytovaném např. předmětu **AD0B36PR2**.

Plán přednášek

1. P: Introduction, Big Data processing motivation, requirements
2. C: Cloud computing cluster OpenStack basic commands, virtualization.
3. P: Hadoop overview - all components and how they work together
    a. **Hadoop Common**: The common utilities that support the other Hadoop modules.
    b. **Hadoop Distributed File System (HDFS™)**: A distributed file system that provides high-throughput access to application data.
    c. **Hadoop YARN**: A framework for job scheduling and cluster resource management.
    d. **Hadoop MapReduce**: A YARN-based system for parallel processing of large data sets.
4. C: install hadoop, hw requirements, sw requirements, how to administer (create access), introduce to the basic setup on our cluster, how to monitor. Run the words histogram, single thread.
5. P: introduction to map reduce, how to use preinstalled data. Basic skeleton for running words histogram Java
6. C: The bag of words notion, TF-IDF, run SVD, LDA.
7. P: HDFS, NoSQL databases, HBase, Cassandra, SQL access, Hive,
8. C: Manipulation with data, how to upscale-downscale HDFS, How to run and monitor computation progres, how to orgnaize the computation.
9. P: What is **Mahout**, what are the basic algorithms
10. C: Run random forest classification task using the Mahout algorithms, show how much faster is the map reduce implementation compared to single thread on one box.
11. P: Streamed data - real time processing
12. Twitter data processing, simple sentiment algorithm
13. Prezentace semestrálních prací a zápočet (IBM)

Plán cvičení

sest cviceni
Cvičení by měla probíhat standardním způsobem předpokládáme, že studenti si přinesou vlastní počítače pro editování skriptů. Vlastní výpočty plánujeme v počítačovém klastru, který nám pro účely výuky propůjčuje centrum.cz. Na cvičeních nebude opakována/probírána látka z přednášek; náplní cvičení bude praktické uplatnění přednášených technologií na konkrétních

příkladech. V průběhu semestru jsou plánovány dva testy z dosud probrané látky.

1. Návrh sémantické HTML5 stránky s ohledem na nové HTML značky
2. Řešení cross-browser compatibility, progressive enhancement, graceful degradation. Zápočet

Zakončení předmětu

Předmět je zakončen klasifikovaným zápočtem. KZ se uděluje na základě individuálního vypracování a obhajoby semestrální práce. Tématem semestální práce bude návrh algoritmu pro zpracování velkých dat technologií hadoop.

Hodnocení studentů

Za předmět je možné získat maximálně 50 bodů:

Aktivita	**Maximum**	**Minimum**
1. test	10	5
2. test	10	5
Semestrální práce	30	15

Výsledná známka odpovídá klasifikační stupnici ECTS, tj. 50-45 výborně, < 25 nedostatečně.

Kapacita předmětu

Kapacita předmětu je maximálně 40 studentů.

Odpovědné osoby

Přednášející: Jan Šedivý
Cvičící: Tomáš Vondra
Cvičící: Tomáš Tunys