TDT4305 Big Data-arkitektur våren 2022
(Førebels faginfo)

Faglærar: Kjetil Nørvåg, noervaag@ntnu.no
Øvingsleiar: Hassan Abedi (hassan.abedi@ntnu.no) (kontaktperson for alt ang. prosjekt)
Forelesning/Q&A-sesjonar: Video/Zoom, på torsdagar 1015-1200
Eksamensdato: Sannsynlegvis 10. juni

Undervisingsopplegg: For å kunne ha eit opplegg som er mest mogleg robust mht. endringar i smittesituasjon og campus-nærvær, vil opplegget til dels vere basert på førehandsinnspelte videoar, med spørsmåls-sesjonar relatert til desse. Forfattarane av læreboka har lagt ut svært gode, profesjonelt innspelte, videoar som dekker pensum i boka, så vi kjem til å basere oss på desse, pluss eigne for tema frå artiklar/tilleggsmateriale. De finn videoane (og læreboka) her: http://www.mmds.org/ Her finn de også foilar laga av lærebokforfattarane, vi kjem i tillegg til å legge ut dei vi sjølve har brukt i tidlegare år. Forventningane for spørsmåls-sesjonane er at deltakarar på førehand har sett videoar, lese pensum, og sett på foilane. Med tanke på at sesjonane er tenkt å vere interaktive vil det ikkje verte gjort opptak av desse.
NB! Heri Ramampiaro som foreleser delar av pensum i opptak er ikkje involvert i faget i 2022, så alle spørsmål ang. pensum skal rettast til Kjetil.

Undervisningsplan med tilhøyrande pensum:

Veke	Tema
2	Faginfo, generell intro til Big Data (ingen videoar til denne forelesninga) Forelesning: Kjetil Nørvåg [ElmasriNavathe] [MMDS, 1]
3	Map-Reduce and the New Software Stack Forelesning: Kjetil Nørvåg MMDS: Video 1-6 [HDFS], [DeanGhemawat], [MMDS, 2], [HDG]
4	Spark Forelesning: Kjetil Nørvåg [Zaharia], [LearningSpark]
5	Gjesteforelesning frå Sikt/Uninett Geir Solskinnsbakk og Emil Henry Flakk Tittel: Defensive Security and Big Data
6	Finding similar items Q&A: Kjetil Nørvåg MMDS: Video 1-6 [MMDS, 3]
7	Mining Data Streams Q&A: Kjetil Nørvåg Forelesning: Heri Ramampiaro (opptak) [MMDS, 4]
8	System for straum-data (Storm og AsterixDB)

	Q&A: Kjetil Nørvåg Forelesning: Heri Ramampiaro (opptak) [Storm], [AsterixDB], [AsterixDB2]
9	Gjesteforelesning frå Cognite Tittel: TBD
10	Advertising on the Web Q&A: Kjetil Nørvåg MMDS: Video 1-4 (på video 3 er kun første 7,5 minutt pensum, på video 4 er kun 9:45 og utover pensum) [MMDS, 8]
11	(ingen organisert aktivitet)
12	Recommendation systems Q&A: Kjetil Nørvåg Forelesning: Heri Ramampiaro (opptak) Recommender systems Del I og II [MMDS, 9]
13	Mining Social-Network Graphs Q&A: Kjetil Nørvåg Forelesning: Heri Ramampiaro (opptak) [MMDS, 10]
14	(ingen organisert aktivitet)
15	(Påske)
16	(ingen organisert aktivitet)
17	Project presentations
18	Project presentations

Kommunikasjon med fagstaben: Fortrinnsvis via Piazza, påmelding her: https://piazza.com/ntnu.no/spring2022/tdt4305. Her kan de poste anonymt for dei andre i klassa (namn berre synleg for fagstaben). Vennligst prøv å unngå email med unnatak av "personlege spørsmål". Bruk av Piazza gjer det meir effektivt for oss for spørsmål som fleire lurer på, og også sjansen for svar/tilbakemeldingar frå andre studentar. Vi oppfordrar alle studentar til å delta aktivt på Piazza, også med svar på spørsmål!

Øvingar: Frivillige øvingar som dekker «teoridelen», m/løysingsskisse, er lagt ut på Blackboard. Basert på erfaringar frå 2021 (lav deltaking) prioriterer vi ikkje å bruke ressursar på organisert aktivitet relatert til desse i år, men spørsmål kan stillast på Piazza.

Prosjekt: Prosjektdelen av TDT4305 (som tel 25% av sluttkarakter) er basert på eit mini-prosjekt, der ein skal utvikle ein applikasjon på BigData-rammeverk (i Java, Scala, eller Python), og skrive kort slutt-rapport. Kan gjerast enten individuelt eller i grupper på to personar (vi kjem ikkje til å godta grupper på meir enn to). Vi planlegg publisering av prosjektet veke 6, første innlevering veke 9 (analyse av datasett), og andre innlevering veke 12 (applikasjon og rapport). Presentasjon (i praksis ca. 10 minutt med diskusjon og test-spørsmål relatert til prosjektet, ingen spesiell førebuing nødvendig) planlagt til veke 17 og 18. Vi vil for dei offisielle fristane i veke 9 og 12 ha ein «grace period» på ei ekstra veke, men anbefaler alle som kan å planlegge å levere til rett tid i tilfelle de får problem eller noko uføresett skulle skje.

Gjesteførelesningar: Vi kjem også til å arrangere to gjesteførelesningar i faget, der vi får bedrifter til å presentere korleis dei handsamar problemstillingar rundt Big Data (applikasjonar, system, etc.). Den første vert frå Sikt/Uninett (veke 5), den andre frå Cognite (veke 9).

Eksamen: Sannsynlegvis heimeeksamen, 2 timar, med bokstavkarakter. Tel 75% på totalkarakter. NTNU vil publisere oppdatert informasjon om dette 27. januar.

Gamle eksamensoppgåver: Er lagt ut på Blackboard. NB! Eksamen for 2020 var utan bokstavkarakter og ikkje nødvendigvis representativ for korleis eksamen i 2022 vert (og av same grunn vart det ikkje laga utfyllande løysingsforslag for publisering). Eksamen for 2021 var heime-eksamen med alle hjelpemiddel tillatne.

Pensum

Pensumlitteraturen ligg i fila Pensum.zip på Blackboard, så det er ikkje noko pensum som må kjøpast.

Intro:

- [ElmasriNavathe] *Fundamentals of Database Systems, 7th ed*, Pearson, s. 911-916 (Big Data), Elmasri & Navathe, 2016.

HDFS:

- [HDFS] *The Hadoop Distributed File System*, Schvachko et al., Proc. of MSST, 2010.

MapReduce:

- [DeanGhemawat]*MapReduce Simplified Data Processing on Large Clusters*, (unnateke kap. 5 og appendix A), Dean and Ghemawat, Proc. of OSDI, 2004.
- [HDG] *Hadoop: The Definitive Guide*, s. 19-37 (MapReduce), White, O'Reilly, 2015.
- *Hadoop: The Definitive Guide*, s. 185-201 (How MapReduce Works), White, O'Reilly, 2015. Kursorisk.
- *Hadoop: The Definitive Guide*, s. 268-273 (Join in MapReduce), White, O'Reilly, 2015. Kursorisk.

Spark:

- [Zaharia] *Resilient Distributed Datasets: A Fault-Tolerant Abstraction for In-Memory Cluster Computing*, Zaharia et al., Proc. of NSDI, 2012. (Unnateke sek. 3.2, 5.2, 6-8)
- [LearningSpark] *Learning Spark*, s. 23-60, Karau et al., O'Reilly, 2015

System for straum-data:

- [AsterixDB] *AsterixDB: A Scalable, Open Source BDMS (unntatt Section 5 og 6),* Alsubaiee et al., PVLDB, 2014.
- [AsterixDB2] *Data Ingestion in AsterixDB*, Grove and Carey, Proc. of EDBT, 2015.
- [Storm] *Storm @Twitter*, Toshniwal et al., Proc. of SIGMOD, 2014.

Tema frå Mining Massive Datasets (MMDS):

- *Mining of Massive Datasets (2019),* Leskovec et al., http://www.mmds.org/
  Kap. 1 (kursorisk)
  Kap. 2.0-2.4.3, 2.5.0-2.5.2
  Kap. 3.1-3.4, 3.8 (unntatt 3.8.5)
  Kap. 4 (unntatt 4.5)
  Kap. 8 (unntatt 8.4.5-8.4.6)
  Kap. 9 (unntatt 9.4)
  Kap. 10.1-10.3, 10.7.1.
  NB! Kapittelreferansar over er frå 2019-utgåva, som er lagt ut på Blackboard.

I tillegg er også prosjekt og øvingar pensum.