TDT4305 Big Data-arkitektur
Våren 2022

Dagens tema

- Praktisk informasjon
- Big Data
- System/infrastruktur for Big Data:

  Intro til Hadoop-økosystemet

Hvem er vi?

- Kjetil Nørvåg

- Hassan Abedi (prosjekt)

TDT4305 Big Data-arkitektur

- **Mål med faget:**
  Gi studentene en innføring i hovedaspekt ved Big Data og Big-Data-system. Emnet tar for seg hovedaspekt ved Big Data. Sentrale tema er rammeverk for Big Data-prosessering (MapReduce, Spark, Storm, etc.), datagruvedrift av Big Data, datastrømmer og analyse av tidsserier, anbefalingssystemer, og analyse av sosiale nettverk.
- **Viktige tema innenfor Big Data *ikke* dekket i detalj av dette faget:**
  Datagruvedrift generelt (TDT4300)
  Maskinlæring/prediksjon (TDT4300 og TDT4173)

Forkunnskapskrav

- Anbefalt forkunnskapskrav
  - TDT4145 Datamodellering og databasesystemer, eller tilsvarende (Som forutsetter TDT4100 Objektorientert programmering og TDT4120 Algoritmer og datastrukturer eller tilsvarende)
- Har med hensikt valgt å ikke forutsette at studentene har hatt TDT4300 Datavarehus/datagruvedrift

Undervisningsopplegg

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.

Forelesningsplan
(vil vert oppdatert i faginfo gjennom semesteret)

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]
…	…

Pensum (vil bli oppdatert)

- Hovedsaklig artikler (ingen egnede bøker som passer fagets "visjon" ☹), ligg i zip-fil på Blackboard:
  - *Fundamentals of Database Systems, 7th ed,* Pearson, s. 911-916 (Big Data), Elmasri & Navathe, 2016.
  - *The Hadoop Distributed File System*, Schvachko et al., Proc. of MSST, 2010.
  - *MapReduce Simplified Data Processing on Large Clusters*, (Unntatt kap. 5 og appendix A), Dean and Ghemawat, Proc. of OSDI, 2004.
  - *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.
  - *Resilient Distributed Datasets: A Fault-Tolerant Abstraction for In-Memory Cluster Computing*, Zaharia et al., Proc. of NSDI, 2012. (Unntatt sek. 3.2, 5.2, 6-8)
  - *Learning Spark*, s. 23-60, Karau et al., O'Reilly, 2015
  - *AsterixDB: A Scalable, Open Source BDMS (unntatt Section 5 og 6),* Alsubaiee et al., PVLDB, 2014.
  - *Data Ingestion in AsterixDB,* Grove and Carey, Proc. of EDBT, 2015.
  - *Storm @Twitter*, Toshniwal et al., Proc. of SIGMOD, 2014.
  - *Mining of Massive Datasets*, Leskovec et al., utvalde tema, sjå faginfo for oppdatert liste.
- "NTNU-produserte" foiler og øvinger er også pensum
- NB! Listen over foreløpig, oppdatert info vil dere til enhver tid finne i faginfo på Blackboard

Prosjekt/øvingsopplegg

- Prosjekt som tel 25% av totalkarakter i faget, meir info neste foil ☺
- Frivillige 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
- Test-spørsmål til kap. 4/9/10, laga av Heri: Meint som hjelp medan de les pensum, svara finn de i pensum
- Oppfordrar til bruk av forumet på Piazza (og dei fleste spørsmål finn ein svar på ved å Google litt...☺)
  - Skriv på engelsk om det er spørsmål relatert til prosjektet

Prosjektdelen av TDT4305

- Fokus på ”hands-on” erfaring fra del 1 av faget (”systemdelen”)
- Grupper på *max* to personer
- To-delt:
  - (Del 0: Installering av Spark på egen maskin og gjennomgang av «quick tour»)
  - Del 1: Bli kjent med Spark, bruke Spark til enkel analyse av data
  - Del 2: Utvikle en applikasjon på BigData-rammeverk (Spark)
- Programmeringsspråk: Java, Scala eller Python
- Planlegg publisering av prosjektet ca. veke 6, første innlevering veke 9 (analyse av datasett), og andre innlevering veke 12 (applikasjon og rapport)
- Årets prosjekt blir tilsvarende tidligere prosjekter, eksempel:

2016:

1) Analyse av Foursquare datasett
2) Sentiment-analyse på tweets i Twitter-datasett (positiv/negativ/nøytral, og mer avansert analyse som f.eks. aggregert over byer i USA), med fokus på ytelse og skalerbarhet

2017:

1) Exploration and analysis of AirBnB dataset
2) Sentiment analysis of AirBnB review, important terms, finding similar AirBnB listings

Eksamen

- 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 4-tiimars heime-eksamen med alle hjelpemiddel tillatne
  - Eldre eksamenar er representative tema-messig, men ein kan forvente litt endring i type spørsmål mht. at det også i 2022 sannsynlegvis vert heime-eksamen med alle tilgjengelege hjelpemiddel


BIG
DATA
INFORMATION
STORAGE
LARGE
SETS
MILLION
SOFTWARE
SAN
NAS
SOCIAL
YEARS
RESULTS
DRIVE
REPLICATION
HANDLE
SOURCES
CHALLENGES
LIMITS
COMPLEX
CAPACITY
HAND
ANNOUNCED
PAST
INCLUDE
NETWORKS
HYPOTHESIS
SCIENCE
END
MOST
NEARLY
MAPREDUCE
FUTURE
CRITIQUES
GOVERNMENT
FACTOR
SYSTEMS
PROCESS
MANAGE
ONLINE
DAY
WORLDWIDE
MAKING
INTEGRATION
VOLUME
ONE
OUT
DEVELOPMENT
TECHNOLOGY
MANY
APPROACH
BILLION
NEW
PARALLEL
STATE
CRITIQUE
USE
PEOPLE
TRAFFIC
CURRENTLY
TECHNOLOGIES
MASSIVE
DEFINITION
USING
FRAMEWORK
BETWEEN
STATISTICS
BUSINESS
EFFECTIVE
VISUALIZATION
DATABASES
SIZE
ANALYTICS
PROCESSING
SIMULATIONS
MANAGEMENT
EXABYTES
DETERMINE
WELL
HIGHER
ALGORITHMS
COMPANIES
PRIVACY
INTERNATIONAL
IMPLEMENTATION
BECOMES
SECOND
NATIONAL
INTELLIGENCE
FLOW
INITIATIVE
ECONOMIC
PROJECT
DISTRIBUTED
PER
TERABYTES
MARKET
VARIETY
RELEVANT
ANNUAL
RELATED
TIME
SENSOR
WOULD
NOW
USED
SET
DEPARTMENT
QUERIES
HIGH
USERS
PETABYTES
AMOUNT
REAL
STRUCTURE
RATE
WORK
APPLICATIONS
INFRASTRUCTURE
ANALYSIS
TERMS
PRIVATE
RESEARCH
YEAR
CAPABILITIES
SUCCESSFUL
NEEDED
USES
LEARNING
LESS
HUGE
INTERNET
EVERY
WORLD
VARIOUS
COMPARED
BASED
TIMES
SHARED
APPROACHES
PROCESSES
SEARCH
ARCHITECTURE
PARADIGM
FUNDING
RECORDS
COST
INSIGHT
TOOLS
CENTER

Terminologi

- *Strukturerte data*: veldefinerte felt, t.d. representert i tabellar
- *Ustrukturerte data*: typisk ”av menneske, for menneske”, t.d. tekst-meldingar
- *Semi-strukturerte data*: typisk sjølv-skildrande (med taggar), t.d. XML og JSON
- *Batch-orienter*t: å køyre ein serie med dataprogram utan menneskelig inngripen (dvs. motsett av ”interaktiv”)
- *Nær-sanntid* (near-realtime): kort forseinking mellom når data vert tilgjengeleg og dei vert behandla
- *Sanntid* (realtime): garantert tid for svar frå data vert tilgjengeleg til dei vert behandla
- *Straum-data* (streaming data): kan sjåast på som ”ordna sekvens av instansar”, t.d. sensor-data eller Twitter-meldingar, med høg innfrekvens, og der ein prosesserer data umiddelbart og utan å ha heile sekvensen tilgjengeleg (heller ikkje det som har kome tidlegare)

Big Data: Definisjonar/karakteristikk

"**Big data** is a broad term for data sets so **large** or **complex** that traditional data processing applications (i.e., DBMS) are inadequate for capturing/storing/managing/analyzing. " – *McKinsey*

"***Big data** is high-**volume**, high-**velocity** and high-**variety** information assets that demand **cost-effective, innovative** forms of information processing for **enhanced insight and decision making**." – Gartner*

Dei 3 (og 5 :) V'ane

- *Volume*: enorme mengder data kontinuerlig generert, både brukargenererte og maskingenererte (t.d. sensorar)
  Ikkje i seg sjølv nok til at det er "Big Data"!
- *Velocity*: hastigheit på datagenerering og prosessering, ofte i sanntid, t.d. umiddelbar deteksjon av trendar på Twitter, ofte straum-data
- *Variety*: variasjon i typar data, kombinasjon av strukturerte og ustrukturerte
- *Veracity*: "sannferdigheit", dvs. kvalitet og nøyaktigheit, stort volum kan kompensere for dette
- *Value*: (potensiale for) verdi av data

Viktige teknikkar (fagområde) i Big Data

- Massiv parallellitet
- Datalagring
- Nettverk
- Databasar/spørjingar
- Tungrekning
- Datagruvedrift
- Maskinlæring
- Informasjonsgjenfinning
- Visualisering
- (Personvern)
- …

”Gamle” teknikkar, så kvifor *Big Data* no?

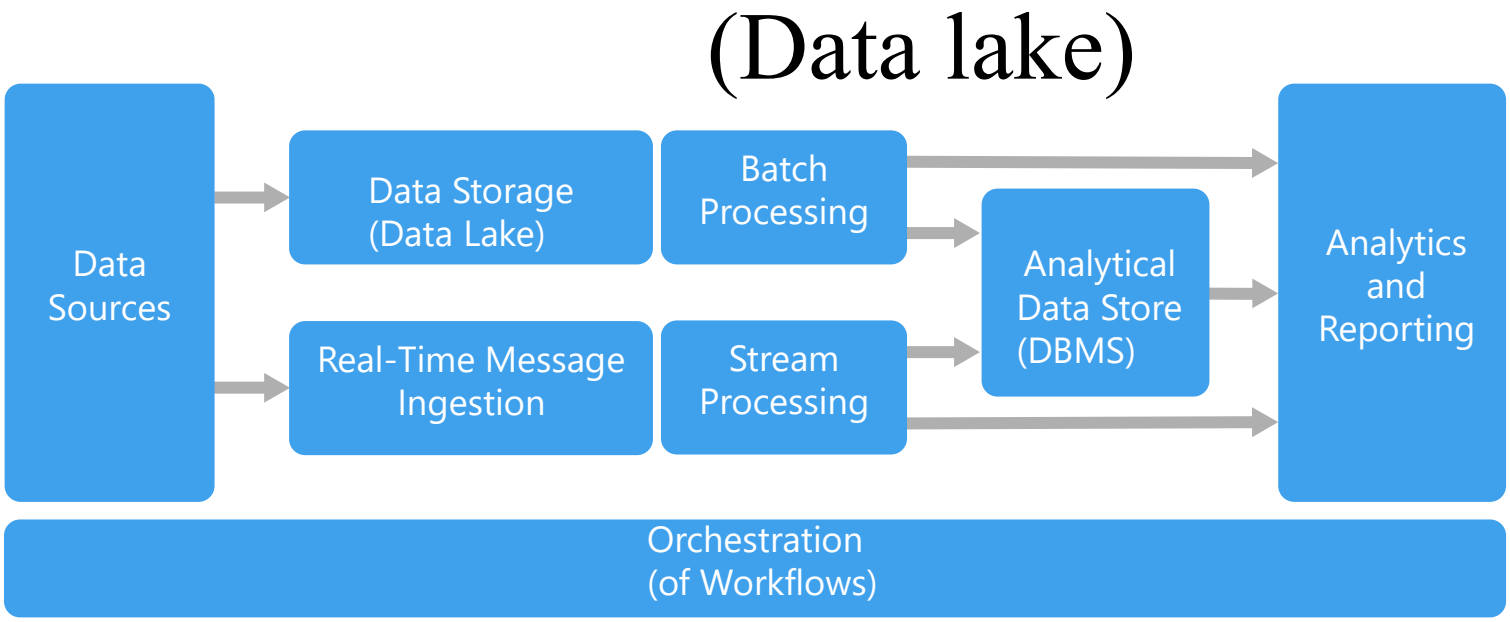
Kvifor Big Data no?

- Meir data innsamla (og lagra):
  - Web (inkl. web-aviser)
  - Sosiale mediar, brukargenererte data
    - Facebook: 1 mrd. kommentarar og 60 mrd Messenger-meldingar pr. dag (april 2016)
    - Twitter: 500 mill tweets per dag (2022), dvs. 6000/minutt!
    - I tillegg: alle data om brukar-aksessar/interaksjon!
  - Transaksjonsdata, t.d. frå telekom
  - Internet of Things (sensorar, RFID, etc.)
- Åpen kjeldekode
  - GNU, Linux, Hadoop, Spark...
- Standard maskin- og programvare (hyllevare)

Eksempel på Big Data-applikasjonsområde (av mange)

- Anbefalingssystem
- Finans-analyse
- Monitorering av sosiale media
- Personaliserte søk
- Trafikk-navigasjon
- Helse (monitorering etc.)

Big Data: Generell arkitektur

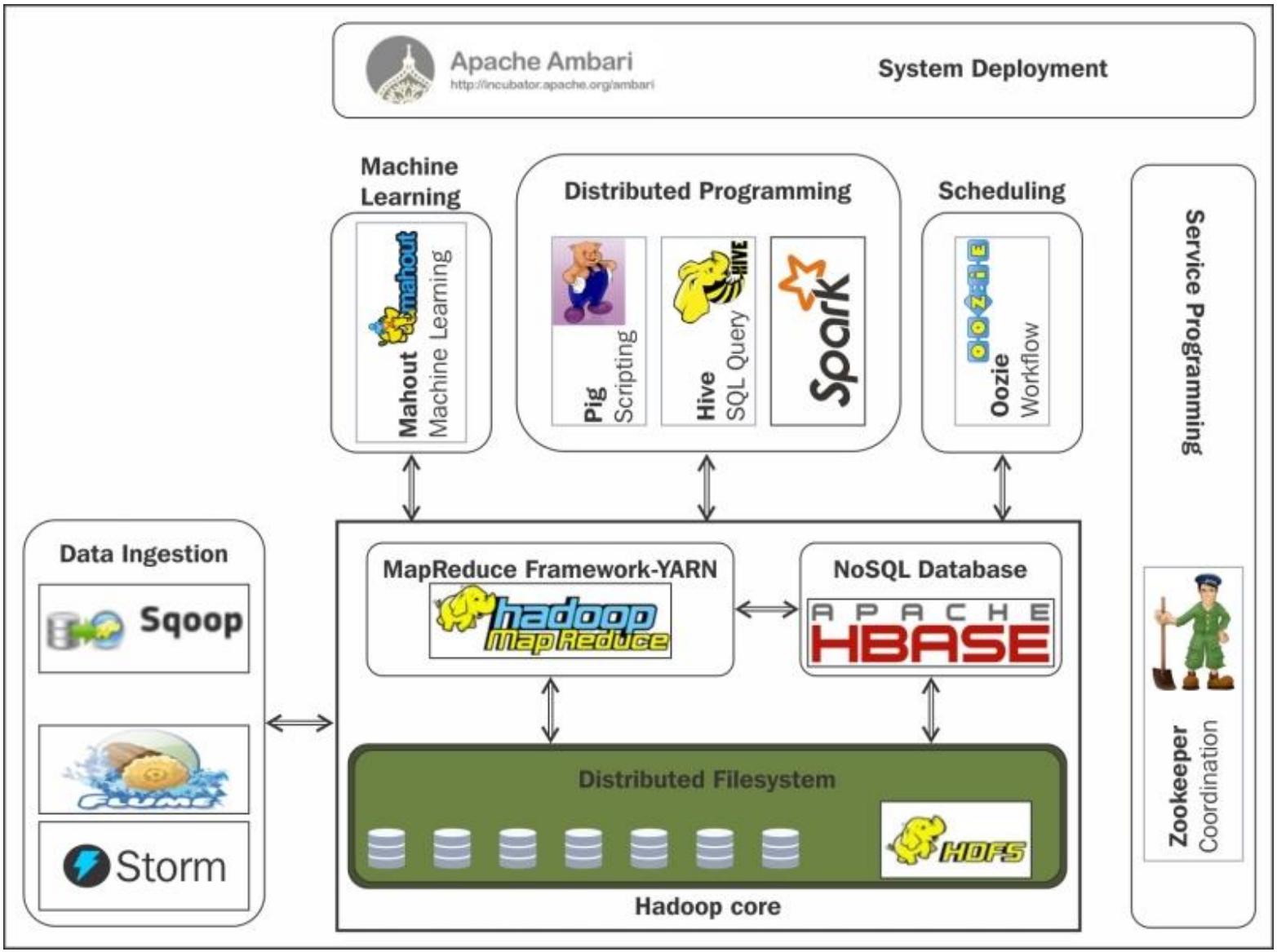


(https://docs.microsoft.com/en-us/azure/architecture/guide/architecture-styles/big-data)

Big Data: Infrastruktur og verktøy

- Hadoop starta i 2006 av Doug Cutting (Yahoo!)
- Verktøy i Apache Hadoop-økosystemet mykje brukt
  - Også som del av kommersielle produkt (t.d. IBM)
  - Hadoop: Distribuert filsystem, ressurs-handsaming (Yarn), prosessering (MapReduce)
  - Mange relaterte Apache-prosjekt som utgjer økosystemet rundt Hadoop
- Apache Spark
- Open-source, men levert i forskjellige ”distribusjonar” (med dyr support ☺) frå kommersielle firma
  - Cloudera, Databricks, pluss «gamle» firma som IBM og HP
- Kan også køyrast i sky-tenestar som t.d. Amazon AWS og Microsoft Azure

Hadoop-økosystemet



(Frå *Hadoop Essentials*)

Interessante datasett fritt tilgjengeleg

- Wikipedia pageviews
- Fullstendig Wikipedia historie
- Sample av tweets vha. Twitter Firehose
  - 1% sample av alle tweets
- Web Science-metode: ”Skrap” nett-stadar sjølv, t.d. allrecipes.com, tripadvisor.com, booking.com etc.