ITFakultetet

Utfordringer & Løsninger

Utfordringer Big Data skaper og hvordan de forsøkes løst

Terje Berg-Hansen

Stavanger 22. mai 2019

ITFakultetet

Før innføring av Big Data

- Analyser kjøres uavhengige av hverandre – i siloer - siden selv nyere skybaserte datavarehus og data science verktøy ikke har blitt designet for å jobbe sammen.
- Data er overalt - i datasentere, offentlige skyer og på enkeltstående servere og pc'er, og det er ingen praktisk måte å kjøre analyser eller maskinlærings-algoritmer på samlede data-sett.
- Med data i siloer og data overalt blir det tilnærmet umulig med en samlet tilnærming til datasikkerhet og privatisering uten å tvinge gjennom kontrollrutiner som begrenser produktiviteten øg øker kostnadene

ITFakultetet

Etter innføring av Big Data

- Data strømmes inn i Data Lakes på klynger av servere og lagres i allment tilgjengelige formater i distribuerte filsystemer, med automatisk replikering og partisjonering («sharding»).
- Data analyseres med parallell-prossessering i klynger av servere – hvor klyngens samlede datakraft utnyttes optimalt – og data fra databaser, filer og strømmer analyseres kombinert i sanntid

ITFakultetet

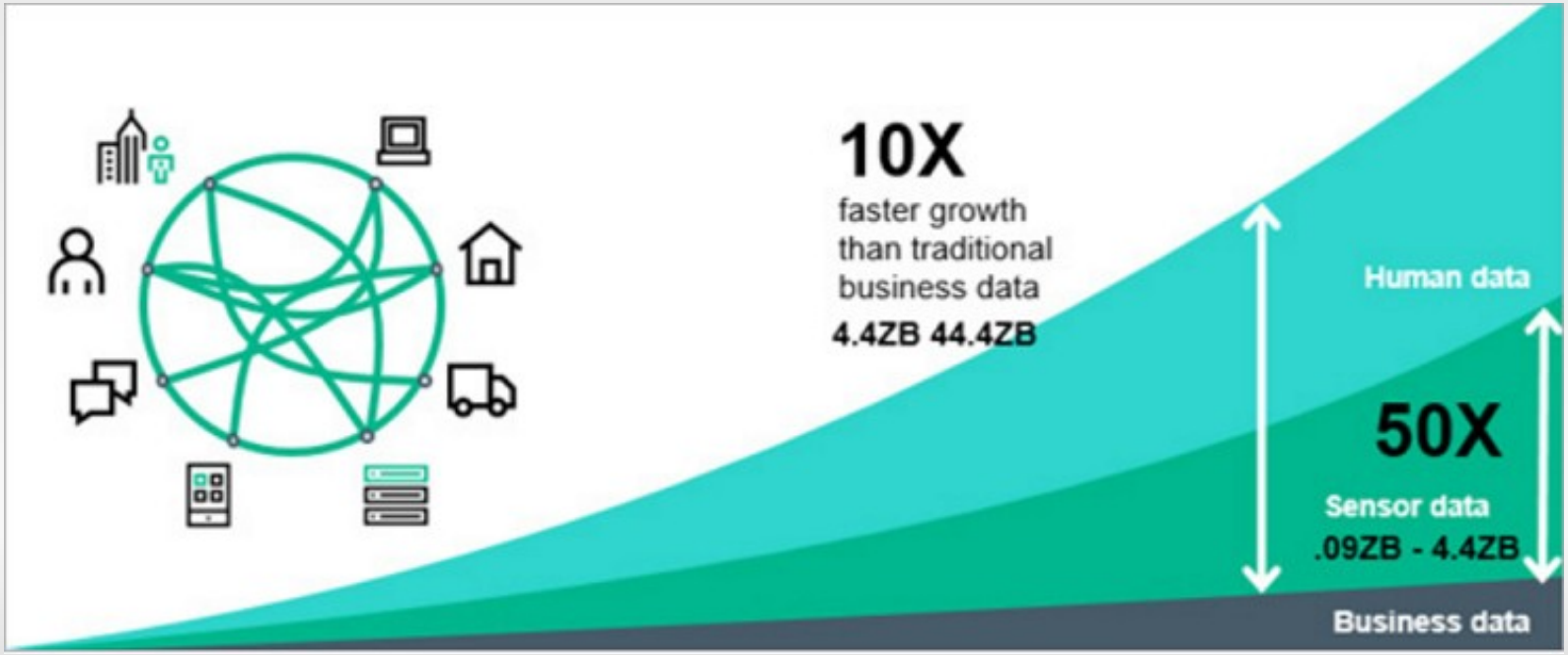
Hovedutfordringer

- Hovedutfordringer knyttet til Big Data
  1) Håndtere økende vekst i datamengde
  2) Bruke data til å gi innsikt på en god måte
  3) Integrere data fra ulike datakilder
  4) Validere data
  5) Sikre data
  6) Rekruttere og beholde relevant kompetanse
  7) Organisasjonell motstand / treghet

ITFakultetet

1) Økende datamengder

- Størrelsen på det digitale universet dobles minst annenhvert år og vil være 50 ganger så stort i 2020 som i 2010.
  - Menneske- og maskin-generert data har 10 ganger så rask vekst og maskingenerert data alene har 50 ganger så rask vekst som tradisjonelle forretningsdata.



ITFakultetet

1) Økende datamengder

Løsninger

- Fra siloer/datatorg til data lakes
- Fra å «skalere opp» til å «skalere ut»
- Partisjonering av data («sharding») på distribuerte filsystemer
- Hybrid-skyløsninger, kombinerer offentlige og sensitive data.
- Analyser utføres med parallell-prosessering i klynger
- Strømming av data for sanntids- eller nær sanntidsanalyser.

Distribuerte Teknologier

- Klynger av servere med:
  - Distribuerte Filsystemer
  - Distribuerte Databaser
  - Distribuerte Analyser
  - Distribuert provisjonering, backup, replikering, monitorering osv.
- Utfordringer bl.a. knyttet til:
  - Systemenes tilstand (state)
  - Lik tid på alle maskiner
  - Riktig Info-spredning i klyngen
  - Partisjonerings-problemer (rack, linjer, rutere, datasentere osv)

ITFakultetet

2) Skape innsikt av data

Utfordringer

- Andelen ustrukturerte og semi-strukturerte data øker
- Datafangst-hastighet og mengde øker
- Behovet for Sanntidsanalyser øker

ITFakultetet

2) Skape innsikt av data

Løsninger

- «Data at rest»
  - Nye verktøy for data-håndtering og analyser bl.a. med schema-fri teknologier
    - Dokument-orienterte og kolonnebaserte databaser som MongoDB, HBase
    - ELT-verktøy som Talend, Spark, Hive
- «Data in motion»
  - Sanntids-analyser via datastrømmer med Real Time Schema-detektering
    - Kafka, Flume, Flink, Spark Streaming,
  - Maskinlæring, Nevrale nettverk (ANN)

ITFakultetet

3) Integrering fra ulike kilder

Utfordringer

- Data kommer fra ulike kilder:
  - Fag-applikasjoner, sosiale media-strømmer, epost-systemer, forretnings-dokumenter etc.
- Det kan by på store utfordringer å kombinere data fra disse kildene på en måte som gjør dem anvendelige for å lage rapporter, analyser osv.
  - Ulike fil-formater, dato-formater, måle-enheter osv.
  - Noen data kan ikke konverteres pga. tekniske eller organisatoriske hindre. (Ref. Ruter / Color Line)

ITFakultetet

3) Integrering fra ulike kilder

Løsninger

- Konvergering mot **åpne standarder**
  - Når «Lego-klossene» er like store er det lettere å bygge sammen konstruksjoner fra ulike leverandører.
- Gode verktøy, systemer og rutiner for integrering
  - ETL => ELT
  - Datakuratering som eget fag / stillingskategori
  - Fleksible systemer som kan håndtere ulike data-formater, samt semi- og ustrukturerte data

ITFakultetet

4) Validering av data

Utfordringer

- Nært beslektet med integrering er validering av data. Ofte får man lignende data fra ulike systemer, og disse dataene er ikke alltid overens om hva som er virkeligheten.
  - Ulike kontaktdata for en kunde, ulike tall fra ulike sensorer for samme måling osv. fører til valideringsproblemer. Hvilken kilde skal være autoritativ?
- «Data Governance» er et begrep som omfatter disse problemstillingene.

ITFakultetet

4) Validering av data

Løsninger

- Utfordringene med «Data Governance» er vanskelige å løse og krever ofte at det opprettes egne grupper som lager regler og policier for å regulere feltet.
- Det fins gode Big Data-teknologier som kan hjelpe, bla. Apache Atlas og Apache Falcon.

ITFakultetet

5) Datasikkerhet

Utfordringer

- Sikkerhet, Brukervennlighet – velg 1
- Det er lett å undervurdere mengden fiendtlig aktivitet og interesse for ens data.
- Det er lett å overvurdere styrken til egne forsvarstiltak.
- Kryptering er kryptisk.
- Skallsikring er ikke tilstrekkelig, men ofte det man har.

ITFakultetet

5) Datasikkerhet

Løsninger

- Etablere identitets- og adgangskontroll, f.eks. gjennom Kerberos kombinert med Ldap/active Directory
- Gi minimums tilgangsrettigheter til filsystem, database o.l. f.eks. med Apache Ranger
- Beskytte web-grensesnitt med f.eks. Apache Knox.
- Gode passordrutiner og kryptering av alt – også epost.

ITFakultetet

6) Kompetanse

Utfordringer

- Behovet for Big Data-kompetanse har drevet lønningene opp for stillinger innen Data Science, AI / Maskinlæring, og Teknisk Drift av Big Data – klynger.
- Utdannelser henger etter teknologiutviklingen
- Det er vanskelig å rekruttere folk med relevant / nødvendig kompetanse

ITFakultetet

6) Kompetanse

Løsninger

- Økt vekt på rekruttering
- Mer intern opplæring av eksisterende ansatte
- Teknologiske løsninger med self-service / innebygd maskinlæring

ITFakultetet

7) Organisasjonelle hindringer

Utfordringer

- Strategien sier datadrevet kultur med Big Data-teknologier, toppledelsen vedtar den, mellomledelsen skal implementere den, uten opplæring eller motivasjon, grunnplanet fikk ikke være med, så da kan de jo være mot.
- IT-avdelingen holder igjen. Big Data var ikke på pensum da de tok sin utdannelse, og det hele virker truende og er antagelig noe konsulent-greier som går over.
- BOHICA (Bend Over, Here It Comes Again)

7) Organisasjonelle hindringer

Løsninger

- Kontinuerlig opplæring av leder, fagfolk og IT-personnel
- Ny stilling: Chief Data Officer
- Nye stillinger: Data Curator, Data Scientist, Data Wrangler osv.
- Ta mellomledernes «skvis» mellom toppledelsen og grunnplanet på alvor og lag strategi/taktikk for hvordan dette skal håndteres.