Høgskolen i Østfold

EKSAMEN

Emnekode:	**Emnenavn:**
ITF301415	Store datamengder: analyse og prosessering Ny/utsatt eksamen
Dato:	**Eksamenstid:**
20.05.2016	09:00-12:00
Hjelpemidler:	**Faglærer:**
Ingen	Edgar Bostrøm Erik Åsberg Davide Roverso
Om eksamensoppgaven og poengberegning:	
Oppgavesettet består av 5 sider inklusive denne forsiden, to sider med oppgaver og to sider vedlegg. Kontroller at oppgaven er komplett før du begynner å besvare spørsmålene.	
Sensurfrist: 13.06.2016 Karakterene er tilgjengelige for studenter på Studentweb senest 2 virkedager etter oppgitt sensurfrist. www.hiof.no/studentweb	

Oppgave 1. Tid 45 minutter.

Gitt følgende tabellstruktur (samme som i øvelsesoppgave gitt i kurset).

STUDENT(studnr, etternavn, fornavn, adresse, ...,postnr) **KURS**(kursnr, kursnavn, nivå)

EKSAMEN (studnr, kursnr, gangnr, dato)

Skriv utsagn i relasjonsalgebra for:

a) Kursnr og kursnavn for kurs på nivå 3.

b) Alt om studenter som har tatt minst ett kurs på nivå 3. *God optimalisering og bruk av semijoin gir best uttelling.*

c) Kurs hvor det ikke er meldt opp noen til eksamen i det hele tatt (i praksis nye kurs som det enda ikke er holdt noen eksamener i). Kursnr og kursnavn på slike kurs skal være med.

d) Studentnr, etternavn og fornavn på studenter som har tatt alle eksamener som det er holdt eksamen i. Tips: prøv først å få til spørringen med bare studentnr på studenter som har tatt alle ....., deretter utvide det med etternavn og fornavn i tillegg.

Oppgave 2. Tid 45 minutter.

a) Beskriv ETL-prosessen i forbindelse med datavarehus.

b) Forklar forskjellen på et «datavarehus» og et «data mart», og forklar «top-down» versus «buttom up» som strategi for oppbygging/utvikling i denne forbindelse.

c) Forklar fordeler og ulemper med bruk av triggere.

d) Hva er cursorer/markører, og hva brukes de til i forbindelse med databaser?

Oppgave 3. Tid 45 minutter.

a) Volume, Variety, Velocity og Veracity brukes ofte til å beskrive et hype't begrep. Hvilket begrep og hva menes med hver av ordene?

b) Forklar prinsippet bak en Key-Value store

c) Hva heter filsystemet som brukes i Apache Hadoop?

d) Hva slags hardware kreves for å kjøre Apache Hadoop?

e) Hvilken markant forskjell er det på Apache Spark og Apache Hadoop?

f) Apache Storm er en teknologi som brukes til et spesielt formål. Hvilket?

Oppgave 4. Tid 45 minutter.

a) Forklar hva som er forskjellen mellom maskinlærings modeller for regresjon og for klassifikasjon. Beskriv et praktisk eksempel for hver av de to.

b) «Overfitting» er et kjent problem når man lager databaserte modeller med maskinlæring. Beskriv hvordan og hvorfor overfitting oppstår og gi noen eksempler av metoder og teknikker man kan bruke for å takle problemet.

c) Forklar kort hva ligger bak begrepet «ensemble model».

VEDLEGG: Relasjonsalgebra - vanlige operasjoner.

Mengdeoperasjoner:	*Notasjon, variant 1*	*Notasjon, variant -2*
Union	$R \cup S$	R union S
Snitt	$R \cap S$	R intersect S
Mengdedifferanse	R - S R \ S	R difference S R minus S
Mengdeprodukt, kartesisk produkt ("alle mot alle")	$R \times S$	R product S R times S
Spesielt for relasjoner:		
Horisontalt utvalg (sigma)	$\sigma_{<betingelse>}(R)$	R where <bet.> R where <bet.>
Vertikalt utvalg (pi)	$\pi_{<attributtliste>}(R)$	R[<attributtliste>]
Mengdedivisjon. (Gitt R[c,d] og S[d]. c er med i mengden R dividert med S hvis c i R forekommer sammen med alle d-er som finnes i S.)	$R \div S$ R / S	R divideby S
Spesialiteter av produkt:		
θ-join (produkt med en eller annen betingelse på kompatible attributter, f.eks. >, <, og komb.)	$R \bowtie_{<betingelse>} S$	R join_{<betingelse>} S (R join S) where <bet.>
Equi-join (θ-opersjonen er =)	" "	" "
Natural join (Equi-join hvor felles attributt kommer bare en gang) ** den mest vanlige jointypen **	" "	" "

Varianter for produkt:		
Outer join, normalt venstre.. (alle i R, samt alle fra S som oppfyller koblingsbetingelsen)	R ⟕ S	R left join$_{<bet.>}$ S
Full join (alle i R, alle i S, samt alle som oppfyller koblingsbet.)	R ⟗ S	R full join$_{<bet.>}$ S
Semijoin (de i R som tilfredsstiller R join$_{<betingelse>}$ S)	R⋉$_{<betingelse>}$ S	R semijoin$_{<bet.>}$ S

Legg merke til at operasjonene her er på mengder, slik at evt. dublikater tas bort – tilsvarende select distinct i SQL.

Dersom betingelsen er på primær/fremmednøkkelkombinasjoner, droppes ofte <betingelse>.