

**Eksamensoppgave i****TDT4145 – Datamodellering og databasesystemer****Mandag 31. mai 2010, kl. 09:00 - 13:00**

*Oppgaven er utarbeidet av faglærer Kjell Bratbergsengen og kvalitetssikrer Svein Erik Bratsberg.*

*Kontaktperson under eksamen er Kjell Bratbergsengen (mobiltelefon 906 17 185)*

*Språkform: Bokmål*

*Tillatte hjelpemidler: D*

*Ingen trykte eller håndskrevne hjelpemidler tillatt.*

*Bestemt, enkel kalkulator tillatt.*

*Sensurfrist: Mandag 21. juni 2010.*

Les oppgaveteksten nøye. Finn ut hva det spørres etter i hver oppgave.

Dersom du mener at opplysninger mangler i en oppgaveformulering gjør kort rede for de antagelser og forutsetninger som du finner det nødvendig å gjøre.

**Oppgave 1 (25 %)**

a) Vi skal lage en database for håndtering av medisinske preparater (medikamenter). Et medikament har et såkalt handelsnavn og en entydig identifikator. Et medikament inneholder ett eller flere virkestoffer. Innholdet av virkestoff i et produkt kan oppgis i mg per g (milligram per gram). Virkestoffet er også identifisert ved et navn og en entydig kode. Et medikament eies av et firma, men kan være produsert av eierfirma eller andre firma (på lisens). Hvis medikamentet er produsert i utlandet er det importert av et importfirma.

Samme medikament finnes i mange forpakninger og styrkegrader. For eksempel kan en salve komme i tuber på 30 gram, 50 gram og 100 gram. Tabletter kan komme i esker eller brett på 20, 50 eller 100 tabletter. Tablettstørrelsen kan også variere, for eksempel 50 mg, 100 mg eller 500 mg.

Et medikament kan erstattes av likeverdige medikamenter av andre produsenter. Databasen skal inneholde opplysninger om hvilke preparater som kan erstatte hverandre. Firma har en egen identifikator og skal registreres ved firmanavn og adresse til hovedkontoret i Norge.

Tegn EER-diagram for en database som kan reflektere opplysningene vi har gitt over. Redegjør for eventuelle antakelser som du gjør.

b) Omform EER-diagrammet til en relasjonsdatabase. Vis hvilke felter som er nøkler og fremmednøkler.

## Oppgave 2 (25 %)

Gitt følgende database for registrering av ulykker og uhell i trafikken, nøkler er understreket og fremmednøkler er skrevet i kursiv og har tabellnavnet de viser til som første ledd:

UHELL (UNr, Dato, Tidspunkt, Vaer\_og\_foereforhold, *ADRESSE\_AdresseId*)

EKVIPASJE (ENr, Dato, *KJORETOY\_KNr*, *PERSON\_Forer\_PNr*)

KJORETOY (KNr, Type, RegNr, Egenvekt, AntPersoner, Farge, Merke)

ADRESSE (AdresseId, Gate\_Vei, *POSTSTED\_PostNr*)

POSTSTED (PostNr, Navn)

PERSON (PNr, ForNavn, EtterNavn, *ADRESSE\_Bor\_AdresseId*)

INNBLANDET (*EKVIPASJE\_ENr*, *UHELL\_UNr*)

PASSASJERER (*EKVIPASJE\_ENr*, *PERSON\_PNr*)

EKVIPASJE sammen med PASSASJERER viser hvilket kjøretøy, fører og eventuelle passasjerer som var "kopleet sammen" på et gitt tidspunkt (da uhellet skjedde).

Svar på følgende oppgaver med SQL-setninger:

a) Skriv ut en oversikt over uhell i året 2009 i Trondheim (Navn i POSTSTED = 'Trondheim'). Skriv ut dato og tidspunkt samt adressen til uhellet.

b) Skriv ut UNr, RegNr, Merke og Farge på kjøretøy samt navnet på eventuell fører som var innblandet i uhellene under punkt a.

c) Lag en oversikt over de personer som har vært innblandet i flest uhell som **fører** av kjøretøy. Tabellen skal sorteres synkende etter antall uhell vedkommende har vært innblandet i. Tabellen skal inneholde personnummer, navn, og antall uhell. Rapporten skal omfatte alle uhell som er registrert i basen.

d) Vis hvordan utførelsen av spørsmålet i oppgave b) kan gjøres ved å sette opp en graf over relasjonsalgebraoperasjoner.

## Oppgave 3 (15 %)

a) Forklar hvordan en randomisert fil (hashed file) er bygget opp. Hva menes med begrepene: hjemmeadresse, synonymer, fyllingsgrad og overløpspost?

b) Forklar hvordan en indeks-sekvensiell fil er bygget opp. Hva menes med begrepene: indeksblokk, datablokk, rotblokk, "fan out", reorganisering?

c) Når bør en bruke indekssekvensiell aksessmetode heller enn randomisert aksessmetode?

#### Oppgave 4 (15 %)

- a) Hva er hensikten med normalisering?
- b) Hvordan er definisjonen på BCNF (Boyce Codd Normal Form)?
- c) Gitt følgende funksjonelle avhengigheter:  $a \rightarrow b$ ,  $b \rightarrow cd$ ,  $g \rightarrow d$ ,  $af \rightarrow c$ ,  $c \rightarrow a$
- i) Finn nøkkelen(e).
- ii) Dekomponer databasen til tabeller på BCNF.

#### Oppgave 5 (20 %)

- a) Hva menes med at en utførelsessekvens er henholdsvis:
- 1) Ikke gjenopprettbar (not recoverable),
  - 2) gjenopprettbar (evt. med galopperende abort (cascading abort)),
  - 3) gjenopprettbar, ikke galopperende abort og
  - 4) strikt?
- b) Gitt følgende historier, i hvilken av klassene under punkt a) tilhører disse:
- H1:  $r1(x)$ ,  $w1(x)$ ,  $r2(x)$ ,  $r2(y)$ ,  $w2(y)$ ,  $c2$ , ....
- H2:  $r1(x)$ ,  $w1(x)$ ,  $r2(x)$ ,  $r2(y)$ ,  $w2(y)$ ,  $c1$ , ....
- H3:  $r1(x)$ ,  $w1(x)$ ,  $r2(y)$ ,  $c1$ ,  $r2(x)$ ,  $w2(x)$ , ....
- c) For at en utførelsesrekkefølge skal være korrekt må den være ekvivalent med en seriell utførelse. Er følgende historie korrekt?
- H:  $r1(x)$ ,  $r1(y)$ ,  $r3(z)$ ,  $r2(y)$ ,  $r3(x)$ ,  $w3(z)$ ,  $w2(y)$ ,  $r1(z)$ ,  $w1(x)$
- Vis hvordan du kommer fra til resultatet.

kjb/18.05.10/

# TDT4145 – Datamodellering og databasesystemer

## Forslag til løsning

Mandag 31. mai 2010, kl. 09:00 - 13:00

### Oppgave 1 (25 %)

a) Vi skal lage en database for håndtering av medisinske preparater (medikamenter). Et medikament har et såkalt handelsnavn og en entydig identifikator. Et medikament inneholder ett eller flere virkestoffer. Innholdet av virkestoff i et produkt kan oppgis i mg per g (milligram per gram). Virkestoffet er også identifisert ved et navn og en entydig kode. Et medikament eies av et firma, men kan være produsert av eierfirma eller andre firma (på lisens). Hvis medikamentet er produsert i utlandet er det importert av et importfirma.

Samme medikament finnes i mange forpakninger og styrkegrader. For eksempel kan en salve komme i tuber på 30 gram, 50 gram og 100 gram. Tabletter kan komme i esker eller brett på 20, 50 eller 100 tabletter. Tablettstørrelsen kan også variere, for eksempel 50 mg, 100 mg eller 500 mg.

Et medikament kan erstattes av likeverdige medikamenter av andre produsenter. Databasen skal inneholde opplysninger om hvilke preparater som kan erstatte hverandre. Firma har en egen identifikator og skal registreres ved firmanavn og adresse til hovedkontoret i Norge.

Tegn EER-diagram for en database som kan reflektere opplysningene vi har gitt over. Redegjør for eventuelle antakelser som du gjør.

### *Løsning*

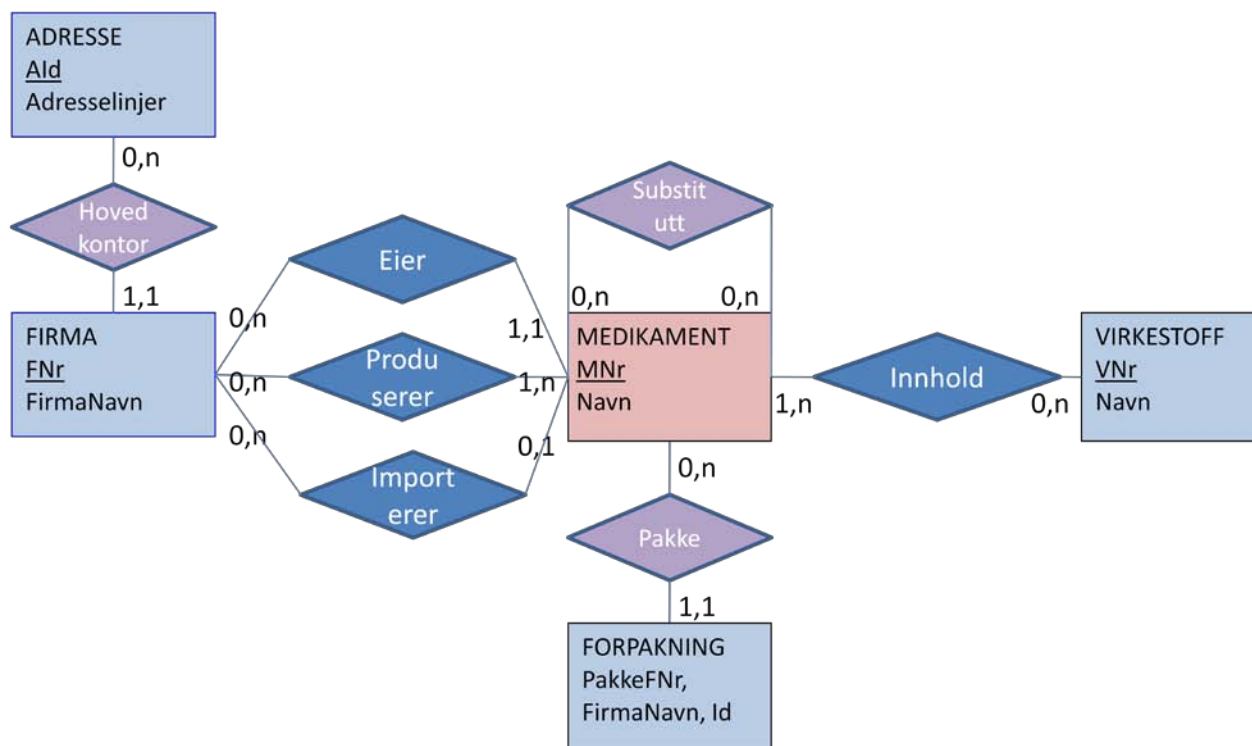
Kommentarer:

Forpakning kan gjerne spesialiseres, i oppgaven er det angitt 2 forskjellige hovedtyper forpakninger: tabletter og salve. Det finnes flere: Flaske eller glass, og sikkert flere som jeg ikke har tilstrekkelig domenekunnskap om. Det var ikke oppgavestillers mening å legge vekt på en spesialisering i denne sammenhengen.

Enheten kan være forskjellig for de forskjellige spesialiseringer. For tabletter vil en kunne angi antall tabletter og tablettvekt, evt. tablettvekt, antall brett og taabletter per brett.

For salve vil det være naturlig å angi størrelsen i gram.

For væske i liter eller ml.



b) Omform EER-diagrammet til en relasjonsdatabase. Vis hvilke felter som er nøkler og fremmednøkler.

MEDIKAMENT(MNr, Navn, *Eier\_FIRMA\_FNr*, *Importør\_FIRMA\_FNr*)

VIRKESTOFF(VNr, Navn)

FORPAKNING(PakkeId, spesialisering, *MEDIKAMENT\_MNr*)

FIRMA (FNr, FirmaNavn, *HovedKontor\_Aid*)

ADRESSE (Aid, Adresselinjer)

Produserer(*FIRMA\_FNr*, *MEDIKAMENT\_MNr*)

Substitutt(*MEDIKAMENT\_MNr a*, *MEDIKAMENT\_MNr b*)

Innhold(*MEDIKAMENT\_MNr*, *VIRKESTOFF\_VNr*, andel)

## Oppgave 2 (20%)

Gitt følgende database for registrering av ulykker og uhell, nøkler er understreket og fremmednøkler er skrevet i kursiv:

UHELL (UNr, Dato, Tidspunkt, Vaer\_og\_foerebeskrivelse, *ADRESSE\_AdresseId*)

EKVIPASJE (ENr, Dato, *KJORETOY\_KNr*, *PERSON\_Forer\_PNr*)

KJORETOY (KNr, Type, RegNr, Egenvekt, AntPersoner, Farge, Merke)

ADRESSE (AdresseId, Gate\_Vei, *POSTSTED\_PostNr*)

POSTSTED(PostNr, Navn)

PERSON (PNr, ForNavn, EtterNavn, *ADRESSE\_Bor\_AdresseId*)

INNBLANDET(EKVIPASJE\_ENr, *UHELL\_UNr*)

PASSASJERER(EKVIPASJE\_ENr, *PERSON\_PNr*)

Svar på følgende oppgaver med SQL-setninger:

a) Skriv ut en oversikt over uhell i året 2009 i Trondheim (Navn i POSTSTED = 'Trondheim'). Skriv ut dato og tidspunkt samt adressen.

```
SELECT u.Dato, u.Tidspunkt, a.Gate_Vei, a.PostNr, p.Navn
FROM ADRESSE a, POSTSTED p, UHELL u
WHERE p.Navn = 'Trondheim'
AND p.PostNr = a.POSTSTED_PostNr
AND a.AdresseId = u.ADRESSE_AdresseId
AND u.Dato BETWEEN DATE '2009-1-1' AND DATE '2009-12-31';
```

b) Skriv ut UNr, RegNr, Merke og Farge på kjøretøy samt navnet på eventuell fører som var innblandet i uhellene under punkt a).

*Løsning:*

Vi lager først et view for å lette lesbarheten. Løsning uten view er like bra.

```
CREATE VIEW TrondheimsUhell AS
SELECT u.UNr
FROM ADRESSE a, POSTSTED p, UHELL u
WHERE p.Navn = 'Trondheim'
AND p.PostNr = a.POSTSTED_PostNr
AND a.AdresseId = u.ADRESSE_AdresseId
AND u.Dato BETWEEN DATE '2009-1-1' AND DATE '2009-12-31';
```

```
SELECT t.UNr, k.RegNr, .Merke, k.Farge, p.ForNavn, p.Etternavn
FROM TrondheimsUhell t, KJORETOY k, EKVIPASJE e, PERSON p, INNBLANDET i
WHERE t.UNr = i.UHELL_UNr
AND i.EKVIPASJE_ENr = e.ENr
AND e.KJORETOY_KNr = k.KNr
AND e.PERSON_Forer_PNr = p.PNr;
```

Denne løsningen har den begrensning at hvis et kjøretøy er uten fører så vil det ikke lages en resultatlinje for dette kjøretøyet selv om det var innblandet i uhellet. For å unngå denne mangelen må vi modifisere siste linjen i selectsetningen til:

```
AND (e.PERSON_Forer_PNr = p.PNr OR e.PERSON_Forer_PNr = NULL);
```

c) Lag en oversikt over de personer som har vært innblandet i flest uhell som fører av kjøretøy. Tabellen skal sorteres synkende etter antall uhell vedkommende har vært innblandet i. Tabellen skal inneholde personnummer, navn, og antall uhell. Rapporten skal omfatte alle uhell som er registrert i basen.

*Løsning:*

På eksamen ble det klart at oppgaveteksten var litt uheldig utformet ved at noen la vekt på ... innblandet i *flest* uhell som ... . Betydde det at en bare skulle liste ut de som hadde flest, ikke alle som

hadde uhell? Min mening var å liste ut alle som hadde hatt uhell, de med flest skulle skrives ut først. Dette er klarere uttrykt i den engelske teksten enn i den norske. De som har tolket oppgaven slik at bare de med flest uhell skulle skrives ut, har kanskje fått en ekstra vanskelighet. Det er tatt hensyn til i vurderingen.

```
SELECT p.PNr, p.Fornavn, p.EtterNavn, COUNT(p.PNr)
FROM INNBLANDET i, EKVIPASJE e, PERSON p
WHERE p.PNr = e.PERSON_Forere_PNr
AND e.ENr = i.EKVIPASJE_ENr
AND i.UHELL_UNr IN (SELECT UNr FROM UHELL)
GROUP BY p.PNr, p.Fornavn, p.EtterNavn
ORDER BY 4 DESC;
```

Forslag til løsning hvor en bare plukker ut den eller de som har hatt flest uhell:

```
SELECT p.PNr, p.Fornavn, p.EtterNavn, COUNT(p.PNr) AS AntUhell
FROM PERSON p,
     (SELECT e.Forere_PERSON_PNr AS Forere,
      COUNT(e.PERSON_Forere_PNr) AS AntUhell
      FROM INNBLANDET i, EKVIPASJE e
      WHERE e.ENr = i.EKVIPASJE_ENr
      AND i.UHELL_UNr IN (SELECT UNr FROM UHELL)
      GROUP BY e.PERSON_Forere_PNr) AS t tt
WHERE p.PNr = t.Forere AND t.AntUhell = SELECT MAX(AntUhell) FROM tt;
```

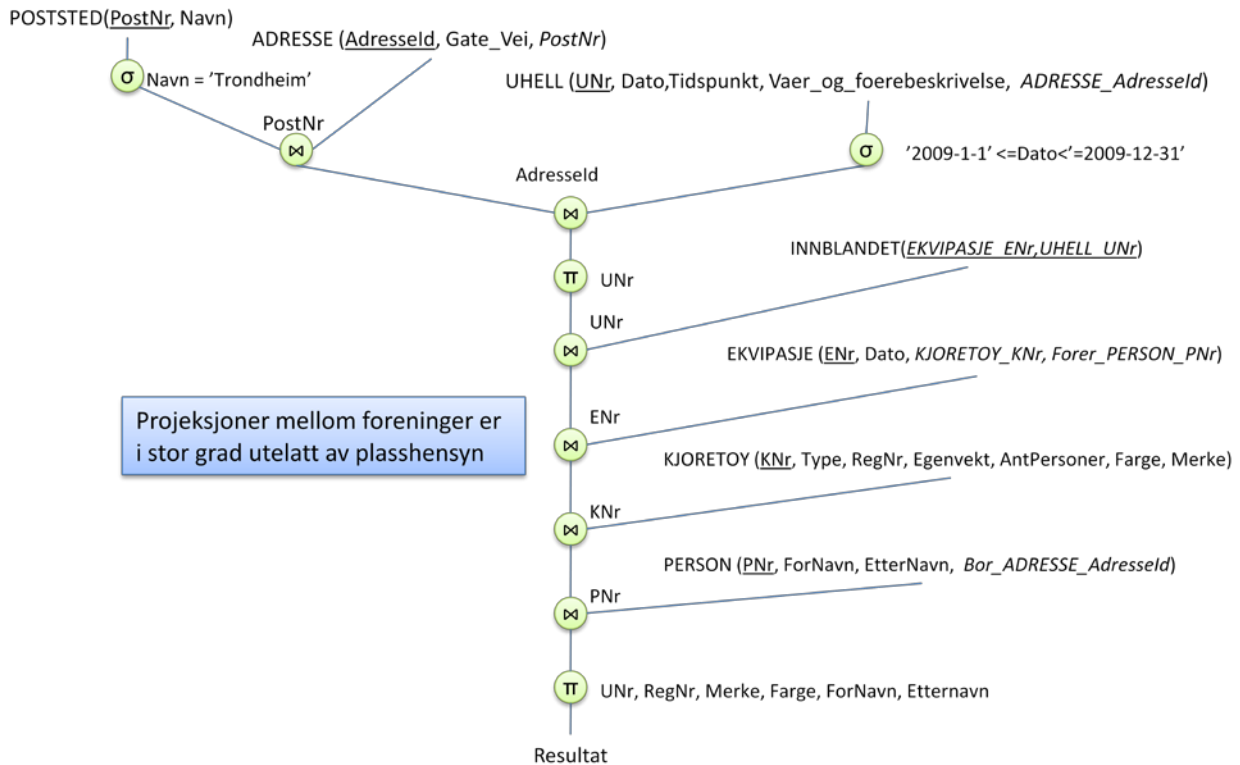
Lager en akkumulert tabell t, tt som inneholder Førere og AntUhell. Setter navn på førere som har maksimalt antall uhell, de skrives ut. Her gir det ingen mening å sortere på antall uhell.

En alternativ måte å løse oppgaven på er å telle antall ganger en person er innblandet i uhell som fører av en ekvipasje:

```
SELECT p.PNr, p.Fornavn, p.EtterNavn, COUNT(SELECT ENr
FROM EKVIPASJE e, INNBLANDET i
WHERE p.PNr = e.PERSON_Forere_PNr
AND e.ENr = i.EKVIPASJE_ENr) AS AntUhell
ORDER BY AntUhell DESC;
```

d) Vis hvordan utførelsen av spørsmålet i oppgave b) kan gjøres ved å sette opp en graf over relasjonsalgebraoperasjoner.

Løsning:



### Oppgave 3 (15%)

- Forklar hvordan en randomisert fil (hashed file) er bygget opp. Hva menes med begrepene: hjemmeadresse, synonymer, fyllingsgrad og overløp?
- Forklar hvordan en indeks-sekvensiell fil er bygget opp? Hva menes med begrepene: datablokk, datablokk, rotblokk, "fan out", reorganisering?
- Når bør en bruke en indekssekvensiell aksessmetode heller enn en randomisert aksessmetode?

Løsning:

- En fil med et fast antall blokker som er adresseområdet. Enten er overløp i egne blokker eller overløpsposter flytter bakover til naboblokker.  
hjemmeadresse – den blokken som adresseformelen beregner for en post.  
synonymer – alle poster med samme adresse tilhører en synonymgruppe eller to poster med samme adresse er synonymer  
fyllingsgrad – det gjennomsnittlige antall poster i hver blokk delt på antall poster per blokk som det er kapasitet til  
overløp – plass for poster som ikke får plass på hjemmeadressen
- En sortert sekvensiell fil med datablokker. Disse kan ligge i fysisk sekvens eller være lenket sammen i en logisk sekvens. For hver datablokk er det en indekspost som inneholder siste nøkkelverdi i datablokken. Indekspostene ligger sortert i en kjede av indeksblokker. For hver indeksblokk er det en indekspost med siste nøkkelverdi i blokken. Disse lagres i en kjede av indeksblokker på nivå 2. Dette mønsteret gjentar seg inntil alle indekser får plass i en blokk – rotblokken.



*indeksblokk – blokk med indeksposter, når datablokkene er kjedet sammen inneholder indeksposten også en peker til blokken.*

*datablokk – den sekvensielle filen med bare dataposter.*

*rotblokk – forklart over.*

*”fan out” – antall indeksposter det er plass til i en indeksblokk.*

*reorganisering – omskriving av hele eller deler av filen for å bedre plassutnyttelsen.*

*B-tre er en indekssekvensiell lagringsstruktur.*

*c) Bruk randomisert lagring ved eksakt nøkkelmatch, indeks-sekvensiell under alle andre forhold (områdesøk).*

#### Oppgave 4 (20%)

a) Hva er hensikten med normalisering?

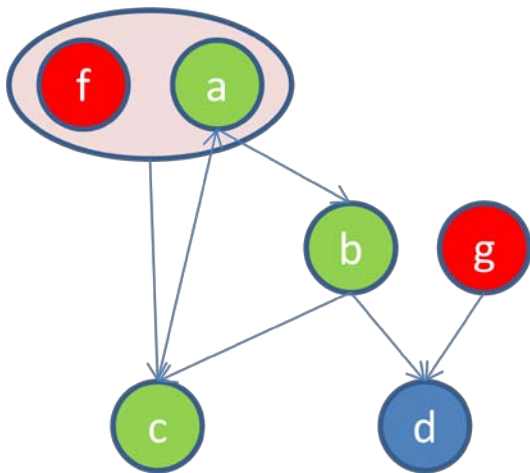
En informasjonsenhet ett sted. Forhindrer anomalier.

b) Hvordan er definisjonen på BCNF (Boyce Codd Normal Form)

*BCNF: Alle determinanter er nøkler*

c) Gitt følgende funksjonelle avhengigheter:  $a \rightarrow b$ ,  $b \rightarrow cd$ ,  $g \rightarrow d$ ,  $af \rightarrow c$ ,  $c \rightarrow a$

i) Finn nøkkelen.

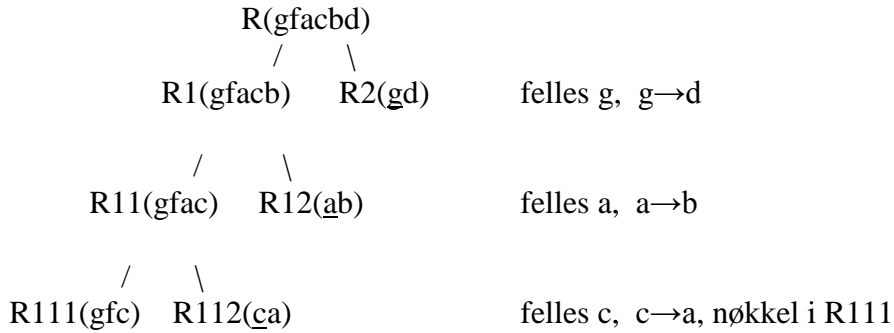


Nøkler er gfa, gfb og gfc. F og g bestemmes ikke av andre enn dem selv og må inngå i alle nøkler. A, b og c bestemmer hverandre og en av dem må inngå i nøkkelen. Dermed får vi tre alternative nøkler.

ii) Lag et skjema på BCNF

For hver (tapsfri) oppdeling må vi passe på at felles attributter i de to nye relasjoner til sammen inneholder alle attributter og at fellesattributtene er nøkkel i en av dem. Fortsetter oppdelingen til *alle determinanter er nøkler* - BCNF

$$R1 \cap R2 = R1 \mid R2$$



$$S = R2(\underline{g}d), R12(\underline{a}b), R112(\underline{c}a), R111(\underline{g}fc)$$

### Oppgave 5 (20%)

a) Hva menes med at en utf\u00f8relsessekvens er henholdsvis:

- 1) Ikke gjenopprettbar (not recoverable),
- 2) gjenopprettbar (evt. med galopperende abort (cascading abort)),
- 3) gjenopprettbar, ikke galopperende abort og
- 4) strikt?

b) Gitt f\u00f8lgende historier, i hvilken av klassene under punkt a) tilh\u00f8rer disse:

H1:  $r1(x), w1(x), r2(x), r2(y), w2(y), c2, \dots$

T1	T2
R(x)	
W(x)	
	R(x)
	R(y)
	W(y)
	c

Ikke gjenopprettbar, det blir feil hvis n\u00e5 T1 skulle gj\u00f8re abort. Da har T2 lest en dirty x og kanskje beregnet ny verdi p\u00e5 y basert p\u00e5 x-verdien.

Da blir det en feil som ikke lar seg gjenopprette.

H2:  $r1(x), w1(x), r2(x), r2(y), w2(y), c1, \dots$

T1	T2
R(x)	
W(x)	
	R(x)
	R(y)
	W(y)
c	

En gjenopprettbar utførelse, men med risiko for galopperende abort. T2 kan nå komittere. Hvis T1 hadde abortert ville også T2 ha måttet abortere.

H3:  $r1(x), w1(x), r2(y), c1, r2(x), w2(x), \dots$

T1	T2
R(x)	
W(x)	
c	
	R(x)
	W(x)

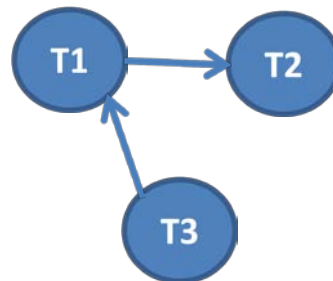
Utførelsen er gjenopprettbare ACA, uten galopperende abort. Kan også sies å være strikt uten at det går klart fram.

c) For at en utførelsesrekkefølge skal være korrekt må den være ekvivalent med en seriell utførelse. Er følgende historie korrekt?

H:  $r1(x), r1(y), r3(z), r2(y), r3(x), w3(z), w2(y), r1(z), w1(x)$

Vis hvordan du kommer fra til resultatet.

T1	T2	T3
r(x)		
r(y)		
		r(z)
	r(y)	
		r(x)
		w(z)
	w(y)	
r(z)		
w(x)		



Konfliktene er mellom operasjoner av samme farge i *forskjellige* transaksjoner. Konfliktgrafene mellom transaksjonene viser ingen sløyfer og utførelsesrekkefølgen så langt er ekvivalent med en seriell utførelse i rekkefølgen T3, T1 og T2. Så på dette stadium kan for eksempel alle transaksjoner komitte.

kjb/18.05.10/