NTNU
Institutt for datateknikk
og informasjonsvitenskap

Eksamensoppgave i

TDT4225 Lagring og behandling av store datamengder

Kontinuasjonseksamen. Fredag 17. august 2012, kl. 0900-1300

Oppgaven er utarbeidet av faglærer Kjell Bratbergsengen.

Kontaktperson under eksamen: Kjell Bratbergsengen, telefon 906 17 185.
NB: Ingen besøksrunder under eksamen, ta kontakt via telefon.

Språkform: Bokmål

Tillatte hjelpemidler: D
Ingen trykte eller håndskrevne hjelpemiddel er tillatt.
Bestemt, enkel kalkulator tillatt.

Sensur: Innen fredag 31. august 2012

Oppgave 1, randomisert lagring (20%)

Det norske folkeregisteret inneholder opplysninger om ca. 8 millioner mennesker. Da regner vi med personer som har eller har hatt et norsk personnummer.
Det norske person-nummeret er bygget opp slik: ddmmyylllcc.
dd er dag i måned, 1-31
mm er måned, 1-12
yy er år, 00-99
lll er et løpenummer, 0-999, liketall for kvinner og oddetall for gutter.
cc er sjekksiffer som beregnes ut fra de andre sifrene i nummeret.

a) Hvilke kvalitetskriterier kan du sette opp for en hashformel?

b) Du skal lage en hashformel som bruker personnummer som nøkkel. Hashformelen skal gi en adresse mellom 0 og 55012101. Begrunn ditt valg av hashformel.

c) Persondata er lagret som personposter på en diskfil som er byte-adressert . Personpostene kommer i vilkårlig rekkefølge og har variabel lengde. Du skal lage en indeks til denne filen ved å bruke hashformelen i foregående oppgave. Indeksen skal ligge på disk. Du skal selv velge hvordan du vil utforme filen. Beskriv indeksfilen.

d) Du skal velge blokkstørrelse, blokkfaktor og fyllingsgrad for indeksfilen. Alle disse valgene skal begrunnes.

Oppgave 2, organisering av data innen en blokk (20%)

Data transporteres mellom arbeidslager (buffer) og disk i blokker.

a) I en blokk lagrer vi administrative data som er til hjelp for å sikre at data blir lagret på en sikker måte. Nevn og forklar noen slike data.

b) En blokk kan inneholde mange poster. Hvordan vil du organisere data i en blokk som skal lagre poster av variabel lengde mellom 20 og 10000 byte? Anta en blokkstørrelse på minst 64 KB og at det er langt flere korte poster en lange. Postene kan slettes og oppdateres. En oppdatering kan endre postlengden – den kan bli mindre like lang eller lengre enn før oppdatering.

c) Du skal lagre poster med følgende felter, se tabellen nedenfor:

Feltnavn	Datatype	Antall byte	Andel (%)
A	integer	4	100
B	char	0-1000	30
C	float	8	90
D	byte	1	100
E	char	30	60
F	char	0-10000	4
G	integer	4	100
H	integer	4	80

`Andel` angir andel i % av postene hvor felttypen forekommer. Felt A er nøkkel i posten. Feltene G og H er fremmednøkler i relasjonsmodellens terminologi. Gjennomsnittslengden av B-felt (som finnes) er 700 byte. Tilsvarende for F-felt er 4000 byte.
Forklar hvordan du vil organisere posten for å oppnå mest mulig effektiv behandling.

d) Hvor lang blir gjennomsnittlig postlengde?

Oppgave 3, Sortering (15 %)

a) Du skal gjøre initiell sortering av følgende poster med reservoarmetoden, reservoaret har plass for 3 poster.
Innfilens nøkler ser slik ut: 32, 44, 10, 3, 62, 79, 64, 43, 98, 33, 8, 19, 5, 2.
Forklar metoden , og vis resultatet ved å skrive ut alle delfilene.

b) Du har fått 432 delfiler og kan flette maksimalt 107 delfiler. Hvor mange ”dummy” delfiler må du legge til i første fletting for å få et optimalt flettetre? Forklar hvordan du kommer fram til svaret

Oppgave 4, 3-dimensjonale rasterbilder (seismikk) (30 %)

3-dimensjonale bilder kan lagres som 3-dimensjonale matriser. Hvert element i matrisen representerer et bilde-element – voksel. Hvert voksel gis en fargekode på 2 byte. Hvert voksel representerer et areal på 10 ganger 10 meter og en dybde på 5 meter når vi lagrer data fra seismiske undersøkelser.

a) Hvor stor datamengde utgjør et seismikk-data for et felt på 20 km ganger 30 km. Det dekkes en dybde fra 0 til 7000 meter.
b) Dataene er lagret på magnetisk disk. Vi velger en blokkstørrelse på 65536 (2^{16})byte. Vi velger å lagre like mange voksler i alle tre akseretninger – x, y og z. z angir dybde. Vi får da en kube eller en likesidet terning med voksler.
   Hvor stor blir sidekanten på terningen som det er plass til i en diskblokk?
c) Data er lagret på disk med følgende parametre: 10000 rotasjoner per minutt, gjennomsnittlig aksesstid er 10 ms og overføringskapasitet er 20 MB/s.
   Hvor mye tid tar det å hente ut en diskblokk?
d) Dataene visualiseres ved å ta utsnitt som er parallelle med ett av akseplanene xz eller yz. Utsnittene er vilkårlig lokalisert. Hvert utsnitt er 1280 x 1024 voksler.
   Hvor stort datavolum må leses for å få fram et slikt utsnitt?
   Forklar også hvordan du resonnerer for å komme fram til svaret.
e) Finn blokkstørrelsen som vil minimalisere *tidsforbruket* for å hente fram et utsnitt etter den metoden som er angitt i oppgave 4d.

Oppgave 5, Relasjonsalgebra (15 %)

a) Forklar relasjonsalgebraoperasjonen divisjon **R=A/B**.

b) Forklar hvordan du vil utføre operasjonen. Skriv gjerne et program i pseudokode.

Kjb/8.8.2012