SUNDHEDSDATA-
STYRELSEN
5. september 2022
FOR – Formidling
Forskerservice

Databehandling på Forskermaskinen

Indhold

Fast servicevindue

Forskerservice kan genstarte alle servere uden yderligere varsel i tidsrummet søndag aften kl. 21:30 til mandag kl. 7:00. En genstart vil betyde, at alle aktive sessioner vil blive afbrudt. Vi anbefaler derfor, at man ikke afvikler jobs søndag aften.

Grundregistre

Forskerservice indgår aftaler med de lokale registeransvarlige i Sundhedsdatastyrelsen om, hvilke registre der kan overføres til Forskermaskinen. Det sker ud fra følgende principper:

- Der skal foreligge hjemmel til, at registeret må udstilles til forsknings- og statistikformål
- Indholdet i registeret skal være dokumenteret
- Brugernes efterspørgsel og behov
- Data skal være rådata med koblingsvariable som CPR-nummer, yder-nummer m.v.

Opdatering af grundregistre

Forskerservice indlæser registre med varierende frekvens på Forskermaskinens SQL-server (databasen), hvis de er opdateret i Sundhedsdatastyrelsens datavarehus.
Opdateringsfrekvensen afhænger af, hvor ofte datakilden opdateres samt mængde og efterspørgsel på data. Det betyder, at nogle registre opdateres ugentligt, andre månedligt og årligt. Opdateringen starter normalt om aftenen.

I overførslen af registrene krypteres identificerende kolonner, som fx CPR-numre og ydernumre. Registrene overføres i samme form, som de ligger i Sundhedsdatastyrelsen datavarehus. Det betyder for en del af de større registre, at de er opdelt i flere tabeller (typisk opdelt efter år).

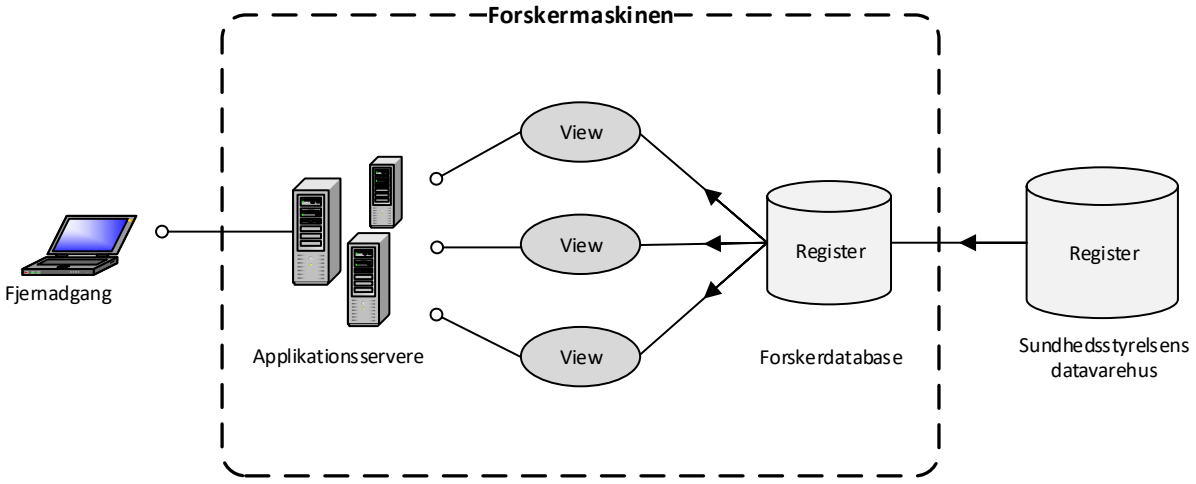
Dataplacering og dataadgang på Forskermaskinen

Registrene på Forskermaskinen er lagret i Forskerdatabasen. Herfra er det muligt at tilgå data med programmerne SAS, Stata og R fra applikationsserverne.

På hvert projekt opretter Forskerservice dataadgang ved enten at danne projektspecifikke views i databasen eller faste datasæt udtrukket fra grunddata. Der gives læseadgang til views og datasæt, som placeres på projektskemaet i databasen.

Vi gør opmærksom på, at det er ikke tilladt at koble data på tværs af projekter. Overtrædelse af ovenstående er brud på Persondataloven.

For views betyder det, at projektet automatisk vil have adgang til den seneste version af registerdata, som er overført fra datavarehuset. For faste datasæt betyder det, at data ikke ændrer sig ved opdateringer af registerdata. Ønskes det at få opdateret data i faste datasæt, skal Forskerservice kontaktes. Den tid Forskerservice bruger på at opdatere data, vil blive afregnet til den aktuelle timetakst.



Figur 1 - Dataflowet på Forskermaskinen. Fastfrosne tabeller udstilles på samme måde som views

Populationsafgrænsning

Når Forskerservice giver dataadgang foretages forskellige afgrænsninger jf. den dataspecifikation, som hører til projektet. Hvis et projekt for eksempel ønsker at forske i en bestemt populations registreringer i LPR, uploades populationen og der dannes views eller faste datasæt til de relevante tabeller, således at det kun er data på de udvalgte CPR-numre, der vil være tilgængelige for projektets brugere.

Et projekts population kan afgrænses på forskellige måder:

- Forsker indsender en population af CPR-numre, hvorefter der gives adgang til populationens oplysninger i udvalgte registre/variable.
- Forskerservice danner en population på baggrund af de i aftalen angivne betingelser, og der gives adgang til populationens oplysninger i udvalgte registre/variable.

Ud over afgrænsningen på populationen kan Forskerservice også afgrænse, hvilke rækker der medtages fra de enkelte registre, samt hvilke variable der medtages. Således kan Forskerservice give adgang til udvalgte records i de enkelte registre for den givne population. I LPR kunne dette for eksempel være ved at udvælge kontakter med bestemte diagnosekoder tilknyttet.

Angående views

Når der gives adgang til registeroplysninger via views på en defineret population, vil registreringerne på populationen løbende ændre sig, når data i databasen opdateres. Dog vil selve populationens størrelse ikke ændre sig, når data i databasen opdateres.

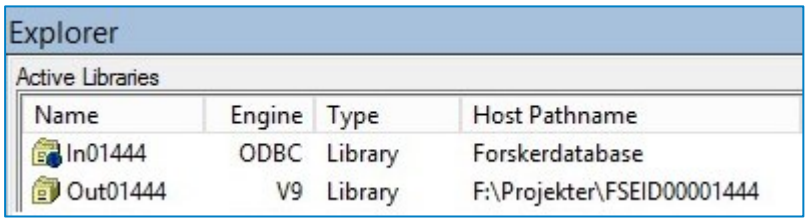
Eksempel
Hvis en view-dataadgang fx består af samtlige data fra LPR på en population bestående af brystkræftramte kvinder, vil der ved hver registeropdatering automatisk komme flere oplysninger om populationen, men populationen vil ikke automatisk blive udvidet med nye kvinder, som er blevet registreret med brystkræft siden sidste opdatering.

Forbindelse mellem applikationsserver og database

Der er etableret en ODBC-forbindelse på alle servere, som brugerne kan anvende. Forbindelsen hedder 'Forskerdatabase'. Det er forskelligt, hvordan man forbinder til databasen fra de forskellige analyseværktøjer. I det følgende har vi beskrevet, hvordan man forbinder med SAS, Stata og R.

Adgang til data via SAS

Når en bruger starter SAS, vil der automatisk blive dannet SAS-libnames, der peger på de skemaer i databasen, som man har adgang til.



Figur 2 – libnames i SAS

Navnet på det pågældende libname starter med IN og efterfølges af projektets nummer (fx IN01444). I libnamet vil brugeren se views og tabeller, der alle fremstår som SAS-tabeller. Man kan ikke skrive til IN-libnamet, da man kun har læserettigheder i databasen.

Man kan selv lave samme libname ved følgende kode:
Libname <navn> odbc src=Forskerdatabase schema=FSEID0000xxxx;
Der vil også blive dannet et output libname til den tilknyttede projektmappe på F-drevet (fx OUT01444). Se eksempel på figur 2.

Her er det også muligt at lave andre libnames til undermapper i projektmappen. Bemærk, at hvis du laver libname til mappen \InputData, så er disse data skrivebeskyttet og kan kun ændres af Forskerservice.

Forbindelse til databasen via SAS SQL Pass-though facility
I takt med at datamængderne vokser, anbefaler vi, at man ved udtræk af større datamængder anvender SAS's mulighed for at sende SQL kode direkte til databasen. Det performer ofte mange gange hurtigere end at anvende SAS/Access libnames.

Her har man således mulighed for at sammensætte, beregne og afgrænse data fra views og tabeller yderligere, inden at data overføres til SAS. Databasen er mange gange hurtigere til at foretage den indledende databearbejdning. Det forudsætter, at man er i stand til at skrive Microsoft SQL (T-SQL) kode direkte mod databasen.

Eksempel – LPR udtræk hvor der anvendes en conncetion til ODBC. Derved kan man skrive T-SQL kode direkte til databasen via SAS.

```
Proc sql;
Connect to odbc (Datasrc=Forskerdatabase);
Create table test as
select * from connection to odbc
/*Her starter T-SQL kode*/
(select CPR_ENCRYPTED
from FSEID0000XXXX.LPR_F_KONTAKTER K
inner join FSEID0000XXXX.LPR_F_DIAGNOSER D
on K.KONTAKT_ID=D.KONTAKT_ID
where DIAGNOSEKODE like 'DN18%' );
Disconnect from odbc;
Quit;
```

Adgang til data via Stata

For at tilgå data på databasen via Stata skal du benytte filen DB Connections FSEID0000XXXX.do, der er placeret i din projektmappe.
Heri er der gemt en række ODBC-kald ned til views og tabeller på dit projekt på forskerdatabasen, som du kan køre for at få adgang til data.

I takt med at datamængderne vokser, så anbefaler vi, at man ved udtræk af større datamængder foretager yderligere sammensætning, beregning og afgrænsning af data. Databasen er mange gange hurtigere til at foretage den indledende databearbejdning. Det forudsætter, at man er i stand til at skrive Microsoft SQL (T-SQL) kode direkte mod databasen.

Det betyder, at du kan erstatte to SQL-sætninger med select * from tabel1 og tabel2 med en SQL-sætning, hvor du joiner og afgrænser data. Hermed kan du spare både kørselstid, RAM på serveren og plads i projektmappen.

Stata anvender RAM, når data bearbejdes. Vi opfordrer derfor alle brugere til at være opmærksomme på de datamængder, der indlæses. Det vil være god programmeringsskik at afgrænse og loope over datamængder, så ressourceforbruget ikke bliver for stort.

Adgang til data via R

Du kan med fordel benytte SAS til at indhente større datasæt, da vi har erfaret, at R har en langsom overførsel fra databasen. Disse kan placeres i projektmappen, hvorefter de kan indlæses i R.

For at tilgå data på forskerdatabasen via R skal du benytte filen DB Connections FSEID0000XXXX.R, der er placeret i din projektmappe.
Heri er der gemt en række eksempler på ODBC-kald ned til views og tabeller på dit projekt på forskerdatabasen, som du kan køre for at få adgang til data. Det er vigtigt, at du ikke afvikler en indlæsning af samtlige tabeller på en gang, da det kan kræve store ressourcer.
Du kan med fordel erstatte select * med SQL-sætninger der joiner, beregner og afgrænser data inden de udtrækkes. Se eksempel fra forrige afsnit.

R og R-studio anvender RAM, når data loades og bearbejdes. Vi opfordrer derfor alle brugere til at være opmærksomme på de datamængder, der indlæses. Det vil være god programmeringsskik at afgrænse og loope over datamængder, så ressourceforbruget ikke bliver for stort.

Herudover opfordrer vi alle R-brugere til at anvende funktionerne rm() og gc() i deres programmering til at sikre oprydning og frigivelse af hukommelse. R anvender stadig RAM, selv om din kørsel er færdig. Det frigøres først, når man rydder op eller lukker session. Hvis alle ressourcerne anvendes på en server, så vil det ikke være muligt for andre brugere at arbejde på den.

Fastfrysning af data

Hvis udtræk af data sker mod views, vil datagrundlaget ændres, når de bagvedliggende registre opdateres. Det kan derfor have den konsekvens, at en reproduktion af resultater ikke er muligt, medmindre data er fastfrossent.

Brugere på projektet skal derfor selv sørge for at gemme en kopi af de nødvendige data. Dette kan i praksis gøres ved at gemme et udtræk af de udvalgte data i projektmappen på F-drevet.

Projektmappen

Alle projekter på Forskermaskinen har en projektmappe, hvor projektets brugere kan arbejde med projektet. Brugerne kan placere analyseresultater, programkode, logs m.m. i projektmappen. Det er desuden muligt at læse data fra projektets skema på forskerdatabasen og gemme data i projektmappen.

Projektmappen har stien: F:\Projekter\<ProjektID>.

Der vil som standard være dannet to undermapper i projektmappen:

- \InputData – Denne mappe anvendes primært af ældre projekter på Forskermaskinen. Mappen er skrivebeskyttet. Som standard bliver dataleverancer placeret på forskerdatabasen og ikke i denne mappe.
- \OutputData – Denne mappe skal anvendes, hvis der skal trækkes data fra projektet ud af Forskermaskinen til behandling på andet databehandlingssted. Udtræk fra Forskermaskinen er kun muligt efter indgåelse af særskilt aftale.

Projektmappen har som udgangspunkt en pladsbegrænsning på 100 GB. Dette kan udvides ved henvendelse til Forskerservice. Vi gør opmærksom på, at du betaler diskforbrug for den plads, du anvender på projektskemaet i databasen og i projektmappen.

Den personlige brugermappe

Brugere på Forskermaskinen har adgang til en brugermappe, som er personlig og kun kan tilgås af brugeren selv. I den personlige brugermappe gemmes personlige indstillinger.

Den personlige brugermappe har stien F:\Brugere\<brugernavn>

Brugermappen skal bruges til at hjemsende aggregerede, anonyme analyseresultater. Mappen kan ligeledes bruges som transfermappe til at dele programkode, der er relevant på flere projekter.

Der må ikke placeres personoplysninger i brugermappen. Det er heller ikke tilladt at flytte data mellem projekter på Forskermaskinen, hverken gennem den personlige brugermappe eller gennem projektmapper. Brugermappen bør heller ikke anvendes til at opbevare øvrige filer, fx programkode eller analyser, der er relevante for projekterne. Disse filer bør placeres i de relevante projektmapper.

Den personlige mappe har en pladsbegrænsning på 100 MB.

Workmappen

Workmappen har stien W:\Brugere\<brugernavn>
Workmappen benyttes til temporære filer fx i SAS, Stata og R. Workmappen kan kun tilgås af den enkelte bruger.

Data bliver automatisk slettet efter 10 dage.

Hjemsendelse af filer

Det er muligt for brugeren at hjemsende de færdige analyser og resultater fra Forskermaskinen. Dette forudsætter, at alle resultater er anonymiseret, så enkeltpersoner ikke kan genkendes i data. Celleværdier med værdierne 1-4 skal i udgangspunktet erstattes af '<5'. Dette gælder uanset datatypen, der hjemsendes (programmer, logs, tabeller og analyser).

Hjemsendelse skal altid ske i henhold til Forskerservices retningslinjer for hjemsendelse af analyseresultater, der er placeret under Forskerservice på Sundhedsdatastyrelsens hjemmeside.

Brugeren skal altid foretage en manuel kontrol af alle filer, inden de hjemsendes. Først herefter kan de flyttes til mailmappen, hvor de automatisk sendes fra. Brugeren må aldrig skrive direkte til mailmappen, da det således ikke er muligt at foretage en kontrol af filerne inden hjemsendelsen.

Mailmappen har stien F:\Brugere\<brugernavn>\mail.

Med ti-minutters intervaller køres der et bagvedliggende script på serverne, der skanner indholdet mailmapperne og sender e-mails af sted til den af brugeren oplyste e-mailadresse med de relevante filer vedhæftet.

Outputs kan kun sendes, hvis filen, F:\Brugere\<brugernavn>\Email.txt, indeholder en e-mailadresse. Når Forskerservice opretter en bruger, vil Email.txt blive oprettet og brugerens e-mailadresse vil placeres heri. Ønsker forskeren at ændre e-mailadressen, skal den rettes i tekstfilen.

Der er begrænsninger på, hvilke filer der bliver sendt fra mailmappen. For billedfiler med filtyperne *.gph, *.png, *.wmf, *.tif, *tiff, *.eps, *.jpg, *jpeg, *.bmp, *.gif er der en størrelsesbegrænsning på 5 MB. For alle andre filtyper er der en størrelsesbegrænsning på 1 MB. Filnavnet kontrolleres også ved hjemsendelse. Ud over tal og bogstaver er den kun tegnene underscore, punktum, runde parenteser, bindestreg og mellemrum, der er tilladt i filnavne.

Behandlede filer fra mappen "\mail" bliver gemt i en af disse mapper:
- Afsendte mails F:\Brugere\<brugernavn>\Sendte mails
- Ikke afsendte mails F:\Brugere\<brugernavn>\Ikke sendte mails.

Det er vigtigt at være opmærksom på, at begrænsningerne på mailfunktionen er oprettet for at sikre, at personfølsomme oplysninger ikke bliver sendt ud fra Forskermaskinen.
Forskerservice tager sikkerhedskopi af alle filer, der forsøges sendt ud fra maskinen. Vi laver løbende stikprøvekontrol af de filer, som er afsendt. Overtrædelser af reglerne kan betyde lukning af data- og brugeradgange for projekt og autoriseret institution.

Er du i tvivl, om du må hjemsende en fil, så henvend dig til Forskerservice. Det samme gælder i tilfælde, hvor du mener, at reglerne kan være overtrådt.

Support og ændringer via henvendelsesformular

Hvis du har behov for support eller ændringer til dit projekt, skal du anvende vores henvendelsesformular som du finder på vores hjemmeside under Igangværende aftale.

Henvendelser kan fx være:
- Problemer med at tilgå Forskermaskinen
- Ønsker til software og opsætning
- Anmodning om adgang til flere data
- Indlæsning af egne data
- Oprettelse eller nedlæggelse af brugeradgange til et projekt

Egne data til projektet

På Forskermaskinen er det muligt at koble registerdata fra en ekstern kilde eller egne indsamlede data til projektet. Dette betegnes som egne data.

Personhenførbare egne data

Ønsker du at få tilføjet egne data til projektet, som indeholder personhenførbare oplysninger, med henblik på, at de skal kobles med registerdata fra Sundhedsdatastyrelsen, skal du være opmærksom på følgende:

- Behandling af dine egne data skal være omfattet af projektets fortegnelse hos din dataansvarlige institution.
- Hvis der er behov for godkendelse fra andre instanser i forbindelse med indsamling af dine egne data (fx fra Styrelsen for Patientsikkerhed), skal denne foreligge, førend data kan indlæses på Forskermaskinen.
- Når du beder om at få egne data indlæst på Forskermaskinen, vil du blive bedt om at udfylde et skema med oplysninger om de data, der indsendes. Oplysningerne om dine egne datasæt indsættes i din aftale om dataadgang. Hvis de indsendte data ikke svarer, til det du har oplyst, kan de ikke indlæses på Forskermaskinen. Det drejer sig om følgende oplysninger:
  - Kilden til data (hvor stammer dine data fra?)
  - Omtrentlig størrelse på datasættet i GB (max. 100 GB per fil)
  - Filtype
    - Dit datasæt skal indsendes som enten en sas7bdat-fil eller en flad fil (.csv eller .txt).
  - Anvendt delimiter ved flad fil
    - Indsender du en flad fil, skal du oplyse, hvilken delimiter, du anvender. Bemærk, at den anvendte delimiter ikke må indgå som værdi i filen. Forskerservice anbefaler brug af ¤ som delimiter, da denne sjældent indgår som værdi i datasæt.
  - Tabelnavn(e)
    - Bemærk, at vi bevarer de angivne tabelnavne, når data indlæses, så navngiv med omhu.
  - For hver tabel skal du angive, hvilke variable der indeholder direkte personhenførbar information, og dermed skal krypteres ved indlæsning. Eksempler på personhenførbare værdier er:
    - CPR-nummer
    - Projektdeltagers løbenummer
    - Patient ID eller unikke prøve ID
    - Ydernumre

    Bemærk, at direkte personhenførbare data, der ikke kan bruges i krypteret form som fx navne og adresser, ikke må indsendes til Forskerservice.

Vær opmærksom på følgende vedrørende navngivning af dine variable:

- Variabelnavnene skal være SAS-kompatible. Vi anbefaler følgende:
  - Vælg variabelnavne, der er max 32 tegn lange
  - Anvend ikke mellemrum i variabelnavnet (anvend evt. _ til angivelse af mellemrum)
  - Anvend ikke specialtegn i variabelnavnet (fx ÆØÅ,.;:`*#%&/)()
- Labels i SAS-datasæt indlæses ikke på Forskermaskinen, navngiv derfor dine variable med omhu.

Bemærk endvidere, at følgende datatyper som udgangspunkt ikke kan indlæses på Forskermaskinen:

- **Lange karaktervariable**

  Antallet af karakterer i en variabel skal være < 40. Eventuelt længere prosa-variable, bør i stedet indsendes grupperet og standardiseret. Hvis du har behov for tekstvariable på > 40 karakterer, til opfyldelse af projektets formål, bedes du fremsende en begrundelse for dette, samt variabelnavn på den pågældende karaktervariabel.

- **Timestamps**
  Hvis dit datasæt indeholder tidsangivelser, der er på time/minut/sekund niveau, skal disse som udgangspunkt reduceres til dato-variable forud for indsendelse. Hvis der er behov for selve timestampet, til opfyldelse af projektets formål, bedes du fremsende en begrundelse for dette, samt variabelnavn på det pågældende timestamp.

Når dit datasæt følger ovenstående guidelines, sikrer du, at data kan indlæses på Forskermaskinen uden problemer, og minimerer den arbejdstid Forskerservice skal bruge på at indlæse filerne. Dermed er du med til at sikre, at det angivne tidsestimat til oprettelse af din dataadgang ikke overskrides.

Ansvar og sanktion

Det er dit ansvar som projektansvarlig inden indsendelse af data, at:

- fjerne alle variable, der indeholder navne eller adresser
- tjekke om der er variable, der 'gemmer på' direkte personhenførbare oplysninger, fx CPR-numre, patient ID, eller navne skjult i en karaktervariabel
- oplyse Forskerservice om alle variable der indeholder direkte personhenførbare oplysninger, og derfor skal krypteres.

Overtrædelse af ovenstående bliver anset som et brud på retningslinjerne for arbejdet på Forskermaskinen. Brud på retningslinjerne vil blive sanktioneret på lige fod med hjemsendelse af mikrodata.

Det godkendte datasæt gemmes på projektet på forskerdatabasen, eller hvis det ønskes kan de placeres i InputData-mappen i projektmappen.
Forskerservice gemmer den rå kopi af de indkomne data i 30 dage efter indlæsningstidspunktet.

Indsendelse af personhenførbare egne data

Vær opmærksom på, at det er et brud på Persondataloven at sende personfølsomme oplysninger over (ikke-sikret) e-mail. Egne data skal derfor sendes til Forskerservice gennem vores upload-løsning. Efter aftale med os udsender vi et link pr. mail til upload-løsningen. Data kan herefter uploades sikkert, da der er tale om en krypteret overførsel og opbevaring af data.

Ikke-personhenførbare egne data

Har du egne data, der *ikke* indeholder personhenførbare-oplysninger, som du gerne vil have indlæst på Forskermaskinen, skal du sende dem gennem vores henvendelsesformular, som du finder på vores hjemmeside under Igangværende aftale. Det er vigtigt, at filerne er gemt i et fladt filformat (f.eks. *.sas, *.do, *.txt), således at Forskerservice kan teste og indlæse dem.
Indlæsning af egne data afregnes til gældende timetakst.

Udtræk fra Forskermaskinen

Ønsker du at få udtræk af rådata med CPR-nummer fra et projekt på Forskermaskinen, kan dette ske ved anmodning gennem vores henvendelsesformular. Dette behandles som en anmodning

om fysisk levering af data, hvilket kan betyde begrænsninger på hvad, og hvor meget data der kan udleveres.

Data skal placeres i undermappen "\OutputData" i projektmappen.

Data leveres enten gennem vores downloadløsning eller kan overføres til Danmarks Statistik via SFTP.

Forskerservice gemmer den afkrypterede kopi af data i 30 dage efter leveringstidspunktet.

Lukning af projekt, brugerkonto og brugeradgange

En brugerkonto kan sættes til inaktiv af Forskerservice. Dette kan ske i følgende tilfælde:

- Forskerservice beslutter, at adgang ophører ved overtrædelse af brugeraftale
- Forskerservice beslutter, at adgang ophører ved inaktivitet i mere end 12 måneder

Når en brugerkonto sættes til inaktiv, er det ikke længere er muligt at logge på Forskermaskinen. Efter at brugerkontoen lukkes på Forskermaskinen, så slettes alle filer i brugermappen.

En brugeradgang til et projekt på Forskermaskinen kan lukkes. Dette kan ske i følgende situationer:

- Brugeren ønsker selv, at adgangen ophører
- Projektansvarlig beslutter, at adgang til projektet ophører

Når en brugeradgang til et projekt lukkes, så vil brugeren ikke længere have adgang til projektmappe for det pågældende projekt.

Et projekt på Forskermaskinen kan lukkes. Dette kan ske i følgende situationer:

- Den dataansvarlige institution beslutter, at adgang ophører
- Godkendelsen til behandling og opbevaring af data fra den dataansvarlige institution udløber
- Det autoriserede forskningsmiljø ophører
- Forskerservice beslutter at lukke adgang ved overtrædelse af regler for brug af Forskermaskinen

Når en dataadgang ophører, slettes alle views og faste datasæt til projektet i databasen og brugeradgange til projektmappe fjernes.

God programmeringsskik

Denne guide skitserer nogle programmeringstekniske metoder, som kan hjælpe dig til at spare tid og serverplads ved at udnytte din allokerede plads på Forskermaskinen.

Serverne har ikke ubegrænset kapacitet. Der er mange brugere, som er afhængige af at serverne er til rådighed. Du bør derfor udvikle og teste kode i en mindre skala, inden du starter en stor kørsel. Gennemtænk, afgræns og test altid din kode. Overvej om bearbejdning af data med fordel kan opdeles mindre portioner.

Begrænsning af variable

De fleste af grundregistre er organiseret i en enkelt tabel med mange variable. Det gælder fx for Dødsårsagsregisteret og Cancerregisteret. I modsætning til disse registre er data i Landspatientregisteret *normaliseret*, så data er spredt ud over flere tabeller.

Når data ikke er normaliseret, arbejder man med et datasæt med mange variable. Hvis du ikke har brug for alle variable, kan du spare plads og øge performance, ved at begrænse dig til de variable, du reelt behøver tidligt i din databearbejdning.

Gode råd til udvikling af programmer

Når du skal arbejde med store datasæt, er det smart at begrænse antallet af observationer i udviklingsfasen. Du kan begrænse antallet af observationer på to måder. Du kan enten gøre det generelt eller ved at begrænse dig til et enkelt programmeringsstep.

SAS - Sådan kan du fx bruge den generelle løsning ved at definere begrænsningen i dine options:

```
/* Her sættes antallet af observationer, der indlæses til 1000 */
option obs=1000;

/* Her indlæses alle observationer. */
option obs=max;
```

Vær opmærksom på, at ovenstående options kan give problemer, når du laver joins på baggrund af store datamængder, da du muligvis ikke får udtrukket nogle rækker. For at undgå disse problemer kan du sætte en begrænsning for det enkelte programmeringsstep.

Da det er muligt at overse denne option, når man har færdigudviklet sit program, er det en god idé at tilføje lidt information om begrænsningen. På denne måde du kan fjerne begrænsningen igen, når du ønsker at arbejde med den fulde datamængde.

SAS - Sådan kan du fx begrænse antallet af observationer under udvikling af programmet:

```
/*****************************************/
/*        KUN 1000 OBSERVATIONER         */
/*****************************************/

proc sql outobs=1000;
create table OBS1000 as
select a.K_RECNUM, a.C_SGH, a.C_AFD, b.*
from from IN999.LPR2_MDL_T_ADM2018 as a
inner join IN999.LPR2_MDL_T_DIAG2018 as b
on a.K_RECNUM = b.V_RECNUM;
quit;
```

Få databasen til at foretage beregninger og afgrænsninger

Det er en mulighed for at sende SQL kode direkte til databasen. Her har man således mulighed for at sammensætte, beregne og afgrænse data fra views og tabeller yderligere, inden at data overføres til SAS, R eller Stata. Databasen er mange gange hurtigere til at foretage den indledende databearbejdning. Det forudsætter, at man er i stand til at skrive Microsoft SQL (T-SQL) kode direkte mod databasen.

Brug af T-SQL i SAS pass though, Stata eller R.
Her kan man afgrænse data i udviklingsfasen ved brug af select top (1000).

T-SQL - Sådan kan du f.eks. begrænse antallet af observationer under udvikling af programmet:

```
Select top (1000) DATO_START, CPR_ENCRYPTED, KOEN
From FSEID0000XXXX.LPR_F_KONTAKTER
```

Begrebsliste

Allokeret pladsforbrug
Tildelt pladskapacitet på server til projekt og bruger

Bruger
En bruger er en databehandler på et projekt. (fsk)brugernavnet består af 9 tegn fskXxxYyy. Brugeren oprettes i SEB-systemet og som Windows-bruger på Forskermaskinen.

Brugeradgang
En bruger tildeles adgang til projektmappe og brugermappe på applikationsserverne og dataadgang til views på Forskerdatabasen

Dataadgang
Er en tildelt adgang til alle views og tabeller på projektet.

Dataspecifikation
En beskrivelse af hvilke registre, tabeller og variable der findes på projektet samt serverplacering, allokeret pladsforbrug og hvilke bruger- og dataadgange, der skal tildeles.

Forskermaskinen
Det miljø, hvor man som bruger får adgang til Forskerdatabasen (omtales som databasen), som indeholder grundregistre og applikationsservere med analysesoftware som R, SAS og Stata.

Grundregistre
De rå registerdata overført fra Sundhedsdatastyrelsens datavarehus. De omfatter en række sundhedsregistre samt CPR-registeret.

Projekt
En ansøgning behandlet af Forskerservice, hvor der findes en godkendt dataspecifikation. Hvert projekt har et unikt ProjektID som er 13 karakterer langt, hvoraf de første fem karakterer består af FSEID, og herefter følger projektnummeret med foranstillede nuller fx FSEID-00000421.

Projektdatabase
En overordnet projektansøgning med et selvstændigt FSEID, der oprettes som et projekt, hvor der er adgang til et bredere dataudsnit, som kan danne udgangspunkt for udtræk til konkrete projekter. Projektdatabaser er omtalt i en særlig vejledning.

View
Er en virtuel tabel, som giver brugeren mulighed for at læse opdateret data fra en grunddatatabel på Forskermaskinen. Forskerservice danner views til alle projektdatabaser, der afgrænser hvilke variable og hvilke rækker, som kan læses fra tabellen på det enkelte projekt. Forskellen på et view og en tabel er, at et view indeholder en definition af en tabel, og at indholdet først dannes, når der laves en forespørgsel mod view'et. Views har den fordel, at data i view'et automatisk bliver

opdateret, når grundregistrene opdateres. Samtidig så undgår man at have mange kopier af de samme data i databasen (redundans), hvilket ville kræve en meget større kapacitet.