Kravsspecifikation

for cloudbaseret løsning til

lagring af forskningsdata

Udarbejdet af arbejdsgruppe nedsat af CIO-forum og DM Ledelses CAB

Oktober 2016

Baggrund

På baggrund af drøftelser i Universiteternes CIO-forum og det nationale samarbejde i DM LedelsesCAB blev der i foråret 2016 nedsat arbejdsgruppe med at det formål at beskrive krav og ønsker til funktioner, universiteterne og deres faglige miljøer har til et system, der kan benyttes til at lagre, dele og synkronisere forskningsdata i skyen.

Arbejdsgruppens medlemmer er udpeget DM Ledelses CAB Arbejdsgruppen og består af:

- Thomas Villum Hansen, Seniorforsker, TEK, DTU CEN (formand, indstillet af eScience komiteen)
- Haakon Lund, Lektor, Det Informationsvidenskabelige Akademi, HUM, Københavns Universitet (indstillet af eScience komiteen)
- Henning Osholm Sørensen, Lektor, Kemisk Institut, NAT, Københavns Universitet (indstillet af eScience komiteen)
- Anne Gerdes, lektor, Institut for Design og Kommunikation, HUM, SDU (indstillet af eScience komiteen)
- Samuel Schmidt, Lektor, Institut for Medicin og Sundhedsteknologi, Sund AAU, (indstillet af eScience komiteen)
- Kaj Søndergaard Laursen, teamleder for AU-IT's team "Generelle Applikationer", AU (indstillet af CIO Forum)
- Jonas Bardino, softwareudvikler, NBI, KU (indstillet af CIO Forum)

Gruppen blev sekretariatsbetjent af Sekretariatschef Gitte Kudsk, DeIC
Fuldmægtig Jonas Holst-Jensen, Danske Universiteter deltog som observatør for CIO Forum

Gruppens kommissorium er vedlagt som bilag 1.

Arbejdsgruppen valgte på det første møde, at tilføje "brug af data" til kommissoriets formål, der herefter lyder "*På baggrund af drøftelser i CIO-Forum og DM LedelsesCAB nedsættes en arbejdsgruppe, der skal forsøge at beskrive hvilke krav og ønsker til funktioner, universiteterne og deres faglige miljøer har til et system, der kan benyttes til at lagre, dele, bruge og synkronisere forskningsdata i skyen.*"

Arbejdsgruppen vedtog samtidig at fokusere på at fremstille en oversigt over krav til et system, og ikke pege på specifikke løsninger, da det vil kræve en dybere indsigt i de enkelte systemer end tidsplanen tillod.

Generel beskrivelse og forudsætninger

Kommissoriets opgave og opstart kunne måske give indtryk af at det handlede om valg mellem Microsofts OneDrive og data.deic.dk. Arbejdet gik dog ambitiøst i en bredere retning, hvor hovedfokus var på at afdække alle konkrete akademiske behov. Først mod slutningen blev der skævet mere til ikke at sætte kravene så højt at det ville udelukke enhver kommerciel løsning.

Arbejdsgruppen har valgt at inddele kravene til et cloud storage system i 11 sektioner. Hvert krav er navngivet, beskrevet samt tildelt en vigtighed betegnet med MR (Minimum Requirement) eller W (Wish) som reflekterer, om arbejdsgruppen mener, at det er et absolut minimum eller blot rart at have. Da der er stor diversitet i hvilke krav de enkelte enheder stiller mht. størrelse og sikkerhed, er der lagt vægt på, at systemet skal være en fælles løsning, som kan tilfredsstille alle.

Undervejs i gruppens arbejde kom det frem, at nogle akademiske behov til datalagring/-brug ikke nødvendigvis kan løses optimalt med kun datasynkronisering. For store filer/datasæt kan en løsning med direkte adgang uden synkronisering f.eks. være mere egnet. Kravene er derfor opdelt i specifikke krav til synkronisering og til direkte filadgang. Såfremt en ren synkroniseringsløsning vælges, bør kravene fra direkte filadgang således stadig adresseres af løsningen.

Arbejdsgruppen har på baggrund af kommissoriet fokuseret på at specificere kravene gældende for forskningsdata*. Der er således ikke taget stilling til eventuelle krav specifikt for universiteternes administrative data.

** Forskningsdata afgrænses til digitale data, der indsamles eller skabes mhp. på forskning - i modsætning til offentlige registre og andre kilder, der også kan gøres til genstand for forskning og videnskabelig analyse.(Definition fra National strategi for forskningsdata management 2015-2018 - https://www.deic.dk/sites/default/files/uploads/PDF/National%20Strategi%20for%20Forskningsdata%20Management%202015-2018.pdf)*

Krav

Krav	Navn	Beskrivelse	Vægtning MR/W
Brugervenlighed			
1	Brugervenligt	Lige så nemt at bruge som Dropbox Løsningen skal tilbyde et intuitivt web interface. Med intuitivt menes f.eks. få klik, få skærmbilleder, få og synlige knapper, "hjælp" til handlinger, genkendelighed i "look and feel", konsistens i anvendelse af menuer og overskuelighed. Den brugervenlige designløsning skal samtidig tilbyde navigationsgenveje, der kan understøtte effektiv afvikling for ekspertbrugeren. Datasynkronisering skal køre gnidningsløst i baggrunden.	MR
Filsynkronisering			
2	Multiple devices	Løsningen skal understøtte synkronisering af samme indhold til flere devices	MR
3	Multiple users	Løsningen skal understøtte at flere brugere kan synkronisere delt indhold	MR
4	Adgang til data uden synkronisering	Brugere skal kunne tilgå synkroniseret data direkte, f.eks. via et web interface.	MR
5	Selektiv synkronisering på alle niveauer (projekter/foldere/enkelt filer)	Brugere skal selv kunne definere hvilke mapper og filer der hentes ned lokalt. Dette skal kunne gøres via web interface eller lokal klient.	MR
6	Optimeret synkronisering	Løsningen kan med fordel bruge en blok-baseret synkronisering, og mulighed for komprimering. Herved kan dataoverførslerne begrænses til kun de ændrede dele af filer og båndbredden i nogle tilfælde udnyttes bedre.	W
7	Mange og store filer	Løsningens skal understøtte synkronisering med både mange (min 30 mio) og store (min 3TB) filer.	MR
Datalagring			
8	Skalerbar plads hos udbyder	Uden øvre begrænsning. Uanset om løsning tilbyder filsynkronisering, direkte filadgang eller begge. Se også punkt 7 og 16	MR
9	Arkiveringsfunktion	Løsningen skal understøtte let adgang til langtidsopbevaring "push-to-archive". Langtidsopbevaring kan evt. være i et andet system. Løsningen til langtidsopbevaring skal kunne garantere, at en skrivebeskyttet kopi af datasæt forefindes i mindst 5 år, evt. i et andet sammenkoblet system.	MR
Fildeling			
10	Lokalt, nationalt og internationalt	Brugeren skal let kunne oprette, styre og nedlægge arbejdsgruppper samt dele filer med samarbejdspartnere, også udenfor akademia.	MR
11	Offentlig deling	Brugeren skal kunne dele filer og foldere via links.	MR

12	Udstilling og publicering af data/frysning af data	Det kan give værdi, at løsningen understøtter sammenhæng mellem data og fx publicerede artikler ved at give mulighed for at linke til blivende frosne datasæt. Jævnfør desuden krav 21b Jf. politiske beslutning vedr. ejerskab til data nederst, fx ifbm fratrædelse.	W
13	Decentral administration ved samarbejde	Forskerne administrerer selv adgang og deling. Man skal ikke forbi it-afdelingen for at kunne oprette personer man skal dele data med. Det bør være nemt for en bruger at se, hvem der har adgang til ens delte data (Dataejer er ansvarlig)	MR
14	Access tracking	Løsningen skal understøtte logning af revisionsspor, der registrerer adgang til og ændring af data	MR
Filadgang			
15	Hastighed	Generel brugeroplevelse som ledende kommercielle løsninger. (evt med decideret minimumskrav (fx > 100 Mbit/s)	MR
16	Mange og store filer	Løsningen skal virke med både mange (min 30 mio) og store (min 3TB) filer	MR
17	Dataindsamling og systemmæssig adgang til data til databehandling	Løsningen skal understøtte dataadgang fra flere typer devices/styresystemer. Dataadgang med flere forskellige effektive protokoller (sftp, rsync over ssh og webDAV).	MR
18	Datahåndtering	Løsningen skal understøtte almindelige datahåndtering (kopiering, merging etc.) via web interface og/eller klient uden at data hentes ned lokalt	MR
19	Dataeksport for samlet organisation	Løsningen skal understøtte mulighed for at tage alle data ud, ved fx skift af systemleverandør (komplet data dump)	MR
Indholdshåndtering og samarbejde			
20	Versionering af data	Minimum X versioner af synkroniserede filer gemmes med henblik på at bruger nemt kan hente ældre version frem igen. Gerne med mulighed for nærmere indstilling af hvilke X versioner der gemmes, såfremt det ikke er alle: f.eks. de X seneste eller én per måned for de seneste X måneder. Versionering skal være mulig for alle filtyper.	MR
21a	Mulighed for metadatering	Systemet kan tilbyde mulighed for metadatering i forbindelse med lagring Det skal være muligt for brugeren at tilpasse skabelon for metadatering, så den følger kravene for det videnskabelige område.	W
21b	Mulighed for metadateringi forbindelse med publicering	Systemet skal understøtte obligatorisk metadatering, når der publiceres. Det skal være muligt for brugeren at tilpasse skabelon	MR

		for metadatering, så den følger kravene for det videnskabelige område	
22	Sikkerhedsklassificering	Brugeren skal kunne klassificere projekter eller foldere (ex personfølsomt/fortroligt/offentligt) ved oprettelse, herunder evt. kunne begrænse tilgængelighed	W
23	Søgemuligheder	Beskrivelse?	W
24	Collaboration værktøjer	Løsningen kan med fordel understøtte Change logs, kommentarer, opgaver. Real-time editering af dokumenter ved flere brugere.	W
Device mangfoldighed			
25	Understøttes på Windows, MacOS, Linux, iOS og Android mobilplatforme	Løsningen skal understøtte læse-/skriveadgang fra alle platforme og synkronisering som minimum fra de tre ikke-mobile platforme. Løsningen skal være browser-uafhængig med webadgang fra alle platformene.	MR
Sikkerhed			
26	Sikker datasikkerhed og dataintegritet	Løsningen leverer høj oppetid (standard/99,8 %) og er godt sikret mod datatab og uautoriseret adgang til data. Kryptering anvendes som standard under transport af data.	MR
27	Back-up og restore	Løsningen skal tilbyde nem og hurtig restore, hvis data går tabt/ødelægges (f.eks. af cryptolocker) Jf desuden punkt 20	MR
28	Antivirus på online data	Det vil give værdi, at løsningen tilbyder central scanning, men hovedansvaret ligger lokalt hos brugere.	W
29	Fleksibilitet i sikkerhedsniveau	Løsningen skal understøtte det nødvendige og tilstrækkelige sikkerhedsniveau i forbindelse med f.eks. anonymiserede data, åbne data, persondata og følsomme data uden at gøre brugen mere besværlig end nødvendig. Det gælder såvel dataopbevaring som datatransport.	MR
30	Personfølsomme data	Hvis en løsning skal være bredt anvendelig af alle, skal det være muligt at lagre personfølsomme data på løsningen. D.v.s. løsningen skal overholde krav fra datatilsynet om bl.a. stærk kryptering ved overførsel og lagring. Link evt blot 'stærk kryptering' til https://www.datatilsynet.dk/offentlig/sikkerhed/staerk-kryptering/ som p.t. specifikt nævner AES256.	MR
31	Brugeridentifikation og adgangskontrol	Åbne systemer, men så vidt muligt med genbrug af eksisterende autentifikationer på universiteterne, så brugere undgår at skulle huske endnu en brugerkonto og kode. Mulighed for eksterne brugere er vigtigt - også uden at disse er oprettet som brugere på danske institutioner.	MR
32	Datalokation efter gældende regler	Safe Harbour/privacy shield Bør afklares juridisk	MR
33	Selective device wipe	Løsningen kan med fordel understøtte fjern-sletning af data på tabte eller på anden vis utilgængelige devices.	W

Management			
34	Central adgang til rapportering	Ledelse og it- afdelingen kan have værdi af at kunne udtrække data om forsøg på uautoriseret adgang til data samt statistiske data om forbrug	MR
35	Data governance	Sammenhæng med krav til datamanagement planer. De enkelte universiteter kan have værdi af at kunne tilbyde Content search og e-discovery på lagrede eller publicerede data.	W
Økonomi			
36	Økonomi	Løsningen skal være konkurrencedygtig, og må ikke belaste den enkelte afdelings økonomi væsentligt. Herunder skal pladskvote og antal brugere kunne skaleres billigt, nemt og uden ekstraudgifter.	MR
Support			
37	Support på løsningen, med en acceptabel responstid.	Support indenfor næstkommende hverdag. Kritiske nedbrud håndteres hurtigere. Se også krav til oppetid.	MR
Overvejelser udenfor arbejdsgruppens mandat			
Politiske beslutninger og kommunikation			
	Ejerskab	Hvem ejer data? Hvad sker der med data ved jobskifte/dødsfald? Der bør være mulighed for enten at overdrage ejerskab eller default adgang til manager/IT-afdeling	?

Bemærkninger fra deltagerne

Arbejdsgruppen besluttede at inkludere et afsnit hvor de enkelte deltagere kan forklare og kommentere på særlige forhold og prioriteter ift kravene. Herunder f.eks. særlige behov for bestemte faggrupper eller institutioner med udgangspunkt i medlemmets tilhørsforhold og baggrund.

Jonas Bardino, KU

Min indgangsvinkel til kommissoriet er hovedsageligt ud fra erfaringer som arkitekt og systemadministrator på KUs ERDA: et private cloud (http://dx.doi.org/10.6028/NIST.SP.800-145) system til netop lagring, deling, brug og synkronisering af forskningsdata.

På KUs bio- og naturvidenskabelige fakultet har vi en væsentlig andel af brugere som arbejder med store filer (snese- eller hundredevis af gigabytes) og - datasæt (adskillige terabytes). I gennemsnit ser vi forbrug på over en terabyte per bruger på ERDA. Arbejde med så store datasæt nødvendiggør ikke bare effektive metoder til at overføre data til/fra lagringsløsningen, men også mulighed for at arbejde på dem i skyen uden først at hente dem ned lokalt. Synkronisering er altså ikke altid en praktisk/tilstrækkelig løsning.

Derudover møder vi jævnligt brugere med 'high performance computing'-behov i deres dataanalyse. Den foregår stort set udelukkende på Linux-systemer, hvorfor en løsning også uproblematisk må kunne bruges derfra.

På KU er der stor fokus på at forskere følger den lokale Research Code of Conduct. Herunder er især funktionalitet til skrive-beskyttet langtidsarkivering af forskningsdata essentiel.

Vi har på KU af flere omgange kigget på OwnCloud som mulig løsning til filkronisering bl.a. i ERDA, men har hver gang forkastet den, især grundet den mildest talt blakkede sikkerhedshistorik både for OwnCloud http://www.cvedetails.com/vulnerability-list/vendor_id-11929/product_id-22262/Owncloud-Owncloud.html
selv og for dens centrale byggesten, PHP (http://www.cvedetails.com/vulnerability-list/vendor_id-74/product_id-128/PHP-PHP.html)
. En serie kritiske sikkerhedshuller ('total compromise of system integrity'!) så sent som sidste efterår, bestyrkede desværre kun mistroen. Så selv *hvis* en OwnCloud-løsning opfyldte alle andre behov, ville den ikke have gang på jord til lagring af vigtige forskningsdata på fakultetet - ja, muligvis på hele KU.

Henning Osholm Sørensen, KU

Jeg deltager i kommissoriet, da jeg er en af brugerne til en cloud baseret løsning for dataopbevaring og har en hvis erfaring i brugen af eksistenerne løsninger, f.eks ERDA. I vores gruppe arbejder bl.a. med undersøgelser af diverse egenskaber af porøse medier. Vi karakteriserer typisk de porøse medier via 3D billeddannelse (røntgentomografi, CT skanninger) opsamlet på store forskningsfaciliteter som synchrotroner, der er lokaliseret rundt om i verden. Ved disse eksperimenter indsamler vi omkring 15 Tb data (2-4 dages målinger). De enkelte datasæt består typisk af 15 Gb rå data og 32 Gb (billede på med 2048^3 punkter) rekonstruerede data (ofte lageret som en fil), men for tidsserier er den samlede dataserie er den samlede datamængde per prøve omkring 1.5 Tb. Tidligere har vi gemt data på eksterneharddiske og fragtet disse hjem fra Japan, Schweitz eller hvor vi nu har lavet eksperimenter. Dette er langtfra nogen sikker måde at overføre eller opbevare vore data på: ofte er det ingen backup, diske kan blive forlagt, da de roterer rundt mellem mange personer osv. På baggrund af de indsamlede 3D billeder udarbejdes forskellige modeller for f.eks væskeflydeegenskaber gennem det porøse medie. Dette betyder en yderligere mangedubling af datamængden (3-20 gange). De indsamlede og afledte data deles ofte af personer internt på KU, men i mange tilfælde også med eksterne personer.

Thomas Hansen, DTU

Jeg er seniorforsker og arbejder til dagligt med store datasæt op til 10 TB og med enkelte filstørrelser i omegnen af 1 TB. I det daglige ser jeg er stort behov for at kunne sende data af denne størrelse til samarbejdspartnere både lokalt på DTU, men også nationalt og internationalt. Desuden er der behov for et arkiveringssystem til disse data. Sidstnævnte behøver ikke at ligge i skyen men kan ligge lokalt.

For tiden er vi ofte nød til at degradere data for at kunne få det ned i størrelse og sende det. Dette er ikke en acceptabel løsning da vigtige detaljer i data kan gå tabt. Hvis det fulde datasæt skal sendes, må vi ofte ty til Fedex eller lignende hvilket er langsomt og ikke særligt interaktivt.