3.5 Hajautetun tiedontallennuksen mahdollisuudet ja ongelmat

Analysoidaan esim. geenitiedon yhdistäminen kansalliseen arkistoon vs. pitäminen erillään

Johdanto: Paljon tallennustilaa vaativia aineistoja, kuten kokonaisia genomisekvenssejä, moniulotteisia tietokonetomografiakuvia sekä patologian digitoituja aineistoja on houkuttelevaa kerätä yhteiseen kansalliseen tietovarastoon arkistoinnin ja yhdistämisen helpottamiseksi.

Ongelmaksi saattaa muodostua tiedonsiirtokapasiteetin rajallisuus, mikäli aineistoa pitää siirtää laskennan vuoksi eri paikkoihin. Toinen ongelma voi olla käyttöoikeuksien hallinta eri teknologioita käyttävien järjestelmien välillä. Siirtoviiveiden aiheuttamaa haittaa voidaan lievittää paikallisella välimuistilla, jossa tiedosta on virallisen arkistokopion lisäksi paikallinen käyttökopio tai toisin päin. Toinen menetelmä on laskennan siirtäminen siihen ympäristöön missä data on, mutta tämä ei ole aina mahdollista. Lisäksi käyttöoikeuksien hallitsemiseksi on olemassa erilaisia kansallisia ratkaisuja, joiden kehittämisessä erityisesti CSC – Tieteen tietotekniikan keskus on kunnostautunut.

	Keskitetty tallennus	**Hajautettu tallennus**
Hyvää	Arkistointi ja lupakäytännöt voidaan hoitaa yhtenäisesti ja käytön valvonta on helppoa.	Tietoturva on helppo toteuttaa rakentaa, jos dataa tuottava analysaattori ja raakatietovarasto ovat samassa hallissa vierekkäin, eikä kopioita datasta ole muualla.
Huonoa	Vaatii tehokkaan verkon ja kaikkien laitteiden tukeman käyttöoikeuksien hallinnan.	Useiden keskusten tuottaman datan analysoiminen vaatii erityistoimenpiteitä, kuten (1) rinnakkaistuvan algoritmin, jolloin data voidaan pitää erillään ja ainoastaan laskenta hajauttaa, tai (2) datan kopioimisen ja yhdistämisen laskentaa varten mikäli menetelmää ei voi rinnakkaistaa.

Esimerkki datasta	Vaadittu tallennustila	Siirtoaika 2M-IT - CSC (*
Korkeatasoinen genomin raakasekvenssi (fastq-muodossa), jossa teknisten lukuvirheiden todennäköisyys on pieni	1 TiB	3h 15min
Perustasoinen genomin raakasekvenssi	250 GiB	50 min
Digipatologian yksittäinen skannattu leike (Panoramic 250-laite)	450 MiB	5 s
Radiologian 3D-keuhkokuva	100 MiB	1.1 s

Taulukko. Datamäärien kasvaessa tiedonsiirto muodostuu aidosti hidastavaksi tekijäksi. Taulukon siirtoajat ja tallennustilat ovat viitteellisiä, mutta kuvaavia. *) Laskennassa on käytetty nopeutta 90MiB/s, joka mitattiin 1 GiB-kokoisen tiedoston SFTP-siirrossa CSC:n cPouta-palvelun ja 2M-IT:n Turussa sijaitsevan datakeskuksen väliltä iltapäivällä syksyllä 2019. Yksittäinenkin hyppy välityspalvelinten yli voi pudottaa nopeuden kymmenesosaan alkuperäisestä.

Keskitetyn tallennuksen ongelma	Mahdollinen ratkaisu
Perinteisen verkkolevynäkymän hitaus etäkäytössä	Käytetään isojen datamäärien tallennukseen sopivia tekniikoita, kuten ”dataämpäreitä” (esim. S3 buckets), joissa mm aikaleimoilta ei vaadita reaaliaikajärjestelmien tarkkuutta (kuten POSIX-levyjärjestelmissä).
Luvitusmekanismien toimivuus erilaisissa legacy-mittalaitteissa, joiden käyttöjärjestelminä voi olla eri vuosikymmenten Windows NT ja Unix-järjestelmiä.	Tallennus paikalliseen varastoon ja synkronointi keskitettyyn palvelimeen. Kiiteäkapasiteettinen verkko (kuten Molekyylilääketieteen instituutin FIMM ja Tieteen tietotekniikan keskuksen CSC välillä).
Poliittiset syyt olla keskittämättä dataa maantieteellisesti esimerkiksi valtakunnan rajojen ulkopuolelle.	Datan pitäminen Suomessa, vaikka se maksaisi enemmän. Datan salaaminen tehokkailla menetelmillä. Suomen kustannustaso ei todennäköisesti ole ongelma, koska kansainväliset IT-jätitkin näkevät Suomen houkuttelevana konesalimaana.
Keskitetyn tallennuksen hyöty	**Perustelu**
Syntyy kannustin yhteismitallistaa eri lähteistä tuleva data	Kaikki data kerätään samaan paikkaa, jolloin eroavaisuudet on helppo havaita
Dataa voidaan analysoida samassa paikassa, jolloin laskentaresurssit ja luvitus voidaan keskittää	Yksikkökustannus syntyy laskentaympäristön kuluista jaettuna käyttöasteella

Erityistosaaminen, kuten bioinformaattisten analyysien tekeminen, voidaan keskittää yhteen organisaatioon – jonka ei tarvitse eikä varsinkaan kannata olla maantieteellisesti keskitetty, vaan se koostuu osista kaikkien datantuottajien operatiivisia tiimejä.	Vaikka data keskitetään, palvelu näyttää käyttäjän kannalta samalta konesalin paikasta riippumatta. Ymmärrys datasta ja mittalaitteen toiminnasta on usein siellä, missä mittauksetkin tehdään