

Selvitys suurteholaskennan hyödyntämisestä toisiokäytössä

Tilaaaja: Sosiaali- ja terveysministeriö (STM) / Findata

Aikataulu: 1.1.2026 - 30.9.2026

Toteutus ja vastuut: Selvitys toteutettiin Findatan koordinoimana yhteistyössä CSC:n ja sidosryhmien kanssa. CSC vastasi työpajasarjan koordinoinnista ja raportin valmistelusta.

Sidosryhmät: HUS, Varha, Tilastokeskus ja Aalto-yliopisto.

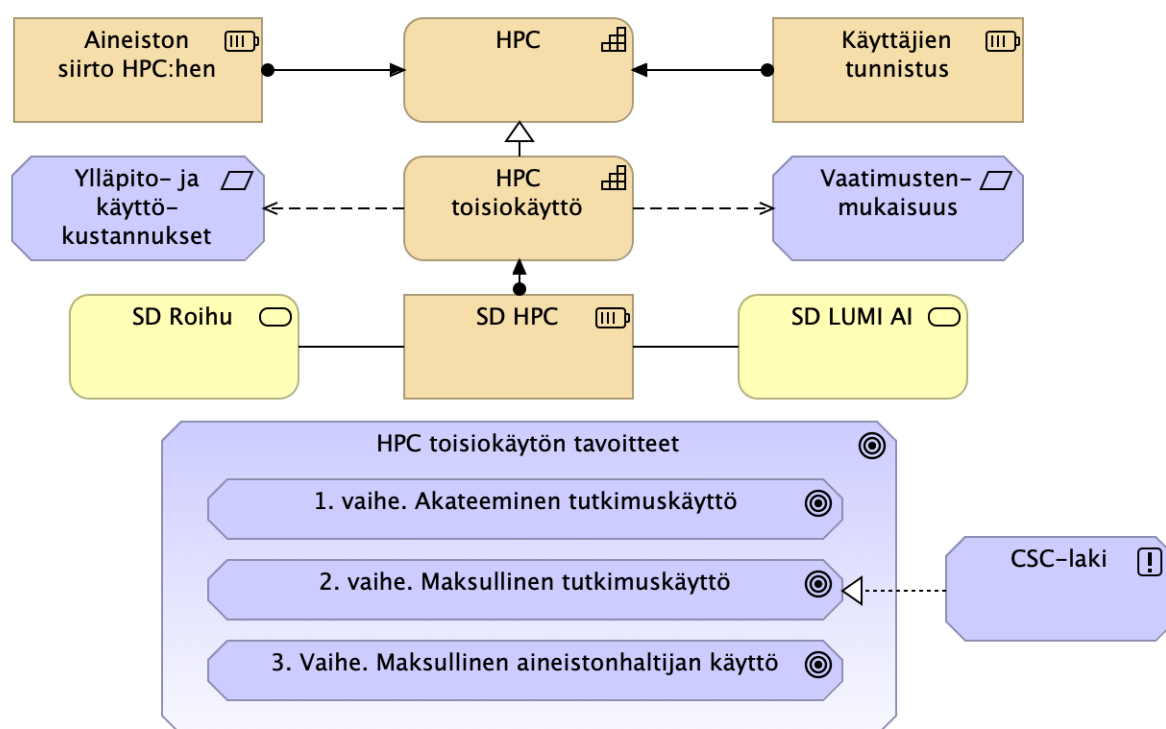
Hankkeen osallistujat:

- Hannu Vilpponen (Findata)
- Maari Parkkinen (Findata)
- Heikki Lanu (Findata)
- Heikki Lehväslaiho (CSC)
- Ilkka Lappalainen (CSC)
- Miikka Kallberg (CSC)
- Jaakko Leinonen (CSC)
- Helena Lodenius (CSC)
- Miika Koskinen (HUS)
- Päivi Lesonen (HUS)
- Juho Kokko (HUS)
- Maria Kauppinen (HUS)
- Arho Virkki (Varha)
- Otto Ettala (Varha)
- Per-Erik Gustafsson (Varha)
- Tommi Saranpää (Aalto)
- Niina Lesonen (Tilastokeskus)
- Ville Koskinen (Tilastokeskus)

1. Yhteenveto

CSC:n suurteholaskennan (HPC) toisiokäytön selvityksessä sidosryhminä toimivat terveystiedon toisiokäytön aineistohaltijoiden edustajat.

Voimassa olevien käytänteiden ja säätelyn erottaessa akateemisen tutkimuskäytön ja terveystietojen käytön toisistaan, toisiokäyttö ehdotetaan aloittamaan tutkimuskäytön mahdollistamisella. Nykyisen toisiolain tarpeiden mukaisen ja laajemman HPC-käytön mahdollistavan kansallisen säännös- ja teknisen valmistelun tulisi tähdätä yhteentoimivaan ekosysteemiin, joka edistää aineistohaltijoiden toimintamahdollisuuksia ja kustannustehokkuutta. Tämä edellyttää CSC:n käyttöoikeuden laajentamista STM:n hallinnonalalle yhteistyössä OKM:n kanssa.



Sisällysluettelo

1. Yhteenveto	3
2. Johdanto	5
Tavoite	5
Selvitystyön rakenne	6
3. Suurteholaskenta	6
CSC:n suurteholaskentainfrastruktuuri ja arkaluonteinen tieto	6
HPC supertietokonejärjestelmän erityisominaisuudet	7
4. Käyttötapaukset ja niiden vaatimukset.....	8
Käyttötapaukset	8
Vaatimukset	9
5. Oikeudellinen yhteentoimivuus	10
Keskeinen sääntely	10
Lailliset reunaehdot CSC:n HPC-ympäristöissä.....	12
6. Organisatorinen yhteentoimivuus	13
Hyvinvointialueet.....	14
Kansallisesti koordinoitu toimintamalli	15
Palveluprosessit ja tulevaisuus.....	17
7. Semanttinen yhteentoimivuus	18
8. Tekninen yhteentoimivuus	19
Toisiokäytön turvavaatimusten toteuttaminen Roihussa	19
Palveluintegraatio	20
Arkaluonteisen tiedon HPC-arkkitehtuuri.....	21
9. Kustannukset ja rahoitusmallit	22
Kustannukset	22
10. Tiekartta ja suositukset.....	23
Liitteet	25
Liite A. Mitä on HPC?	25
Liite B. Käyttötapaukset	27

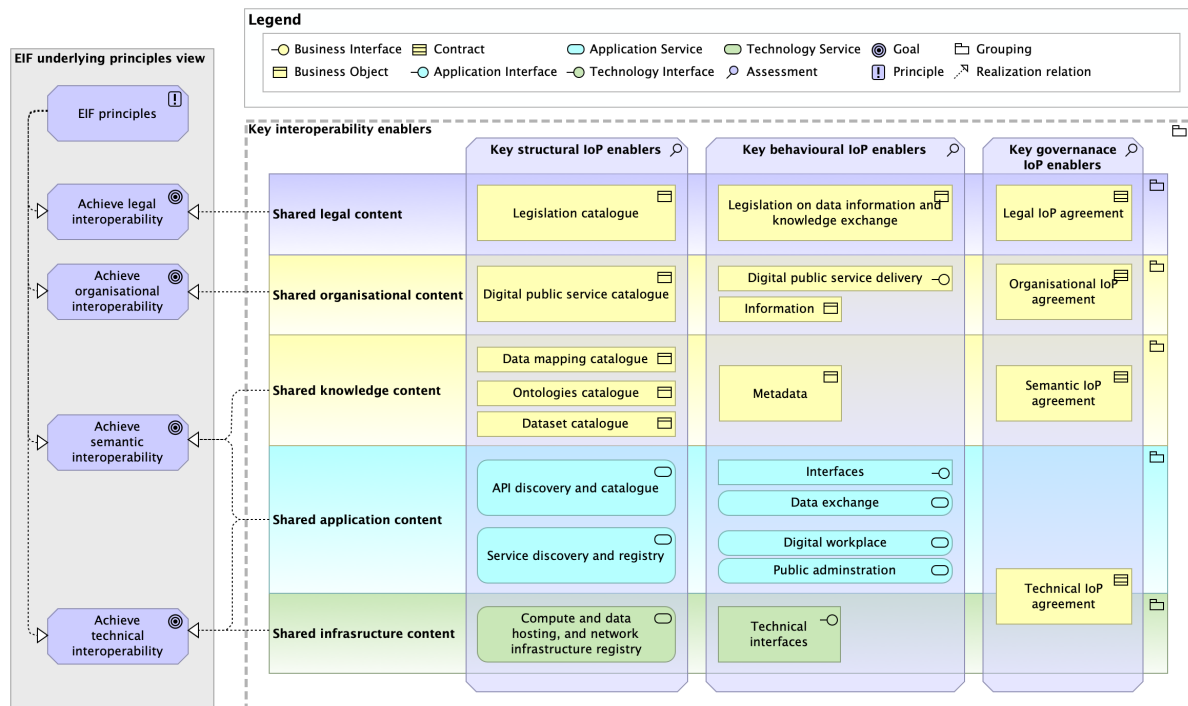
2. Johdanto

Tavoite

Selvitystyön tavoitteena on määritellä, kuinka toisiolain alaisten aineistojen analysoinnissa voidaan hyödyntää CSC:n suurteholaskentainfrastruktuuria. Suurteholaskenta on osa kansallista tietojärjestelmäpalvelujen tavoitearkkitehtuuria, jonka huomioon ottaminen toisiokäytössä voisi johtaa kansallisten resurssien tehokkaampaan ja joustavampaan käyttöön¹. Aineistojen arkaluonteisuudesta johtuva tiukkojen suojaustarpeiden toteuttaminen on kuitenkin haastavaa lähtökohtaisesti vahvasti yhteen kytkeytyssä suurteholaskentaympäristössä.

Selvitystyön pohjaksi kerättiin toisiokäyttöaineistoja tuottavilta terveysalan organisaatioilta laaja kirjo konkreettisia käyttötapauksia, jotka antavat tarpeista riittävän yleiskuvan ja mahdollistavat tarpeiden priorisoinnin. Käyttötapaukset määrittelevät vaatimukset suurteholaskennan ympäristölle, joka liittyy sen muuhun kansalliseen toimintaan.

Tässä raportissa kuvataan palvelutuotannon nykytilanne, haasteet sekä rahoituksen asettamat rajat toteutusehdotuksille. Sen rakenne muodostuu tarkastelemalla kokonaisuuden vaatimaa yhteentoimivuutta neljästä eri näkökulmasta: oikeudellisesta, organisatoriselta, semanttisesta ja teknisestä (Kuva 1).



Kuva 1. Euroopan unionin yhteentoiminnan referenssiarkkitehtuuri EIRA² rakentuu sen periaatteiden neljään pääluokkaan, jotka ovat oikeudellinen, organisatorinen, semanttinen ja tekninen.

¹ Terveystiedon tulevaisuus tekoälyn aikakaudella - Sitra <https://www.sitra.fi/julkaisut/terveystiedon-tulevaisuus-tekoalyn-aikakaudella/>

² EIRA <https://interoperable-europe.ec.europa.eu/collection/european-interoperability-reference-architecture-eira/solution/eira>

Selvitystyön rakenne

Selvitystyö toteutettiin työpajasarjana, johon osallistuivat Findata, CSC, Varha, Tilastokeskus, HUS ja Aalto-yliopisto. Kustakin organisaatiosta kutsuttiin työpajoihin 2–3 edustajaa. Helmi–toukokuun aikana järjestettiin kolme työpajaa. Niissä analysoitiin osallistujien toimittamia käyttötapauksia, tunnistettiin keskeisiä haasteita sekä käytiin keskustelua tavoitetilasta ja HPC-ratkaisun tarpeista.

3. Suurteholaskenta

CSC:n suurteholaskentainfrastruktuuri ja arkaluonteinen tieto

CSC:n suurteholaskentainfrastruktuuri koostuu useasta elinkaarensa eri vaiheissa olevasta superlaskentatietokoneesta. Arkaluonteisen tiedon käsittelylle soveltuvaa HPC-prosessointia, SD HPC-palvelua, on pilotoitu tämän vuoden aikana poistuvassa Puhti-koneessa (Taulukko 1) vuoden ajan käyttäen ei-sensitiivisiä aineistoja. Jo pilotoinnin alkuvaiheessa kävi selväksi, että Puhdin rakenne ei anna arkaluonteiselle tiedolle riittävää suojausta muilta käyttäjiltä.

Nykyiselle LUMI-tietokoneelle tehty selvitystyö vahvisti, ettei sekään pysty täysin varmasti suojaamaan arkaluonteista aineistoa. Sille valmistellaan SD HPC-palvelua, jonka odotetaan palvelevan kaupallisten toimijoiden tarpeita.

Taulukko 1. CSC:n HPC supertietokoneet (kesäkuu 2026 tilanne).

Nimi	Käyttö	Status	CPU-resurssit	GPU-resurssit	Tallennus
LUMI	EuroHPC	Käytössä (EU flagship)	262 000	~24 000 AMD MI250X GCD	117 PB
LUMI AI	EuroHPC	Tulossa 2027			
Mahti	Kansallinen	Poistumassa 2026	180 000	96 Nvidia A100	8.7 PB Lustre
Puhti	Kansallinen	Poistumassa 2026	27 280	320 Nvidia V100	4.8 PB Lustre
Roihu	Kansallinen	Käytössä 2026	188 000	528 Nvidia GH200	6.5 PB Lustre

Nykyinen LUMI-supertietokone korvataan vaiheittain uuden sukupolven EuroHPC-supertietokoneella, LUMI-AI:lla vuoden 2027 aikana. Uuden supertietokoneen laskentateho muodostuu pääosin uuden sukupolven GPU-prosessoreista, jotka soveltuvat hyvin tekoälykehitykseen. Sen teknisistä turvakyvyyksistä ei vielä tiedetä, mutta tavoitteena on tuki joustavaan arkaluoteisen tiedon käsittelyyn.

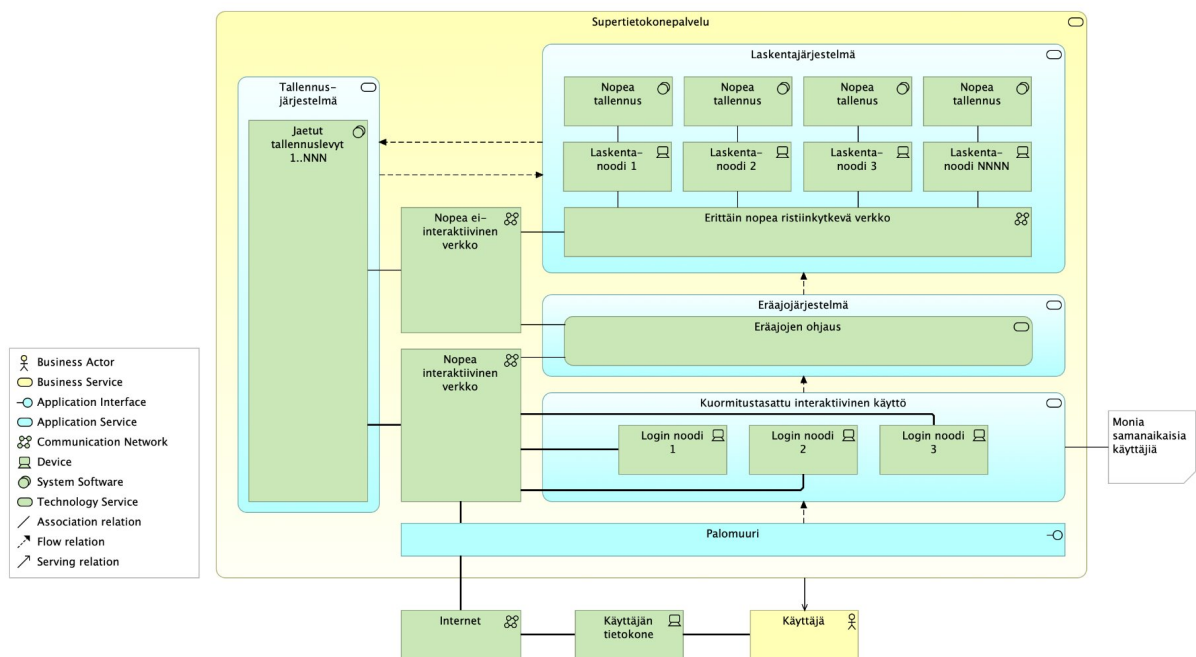
Kansallinen Roihu supertietokone on otettu käyttöön 29.6.2026. Siinä on merkittävästi enemmän GPU-kapasiteettia kuin CSC:n nykyisissä supertietokoneissa Puhdissa ja Mahdissa, ja suorituskykyinen tallennusjärjestelmä mahdollistaa suurten data-aineistojen tehokkaan käsittelyn. Roihun teoreettinen huipputeho on yhteensä 49 Pflops eli 49 tuhatta biljoonaa laskutoimitusta sekunnissa.

HPC supertietokonejärjestelmän erityisominaisuudet

HPC supertietokoneen yleisominaisuudet ja yleisimmät käyttötarkoitukset on kuvattu liitteessä A. Käyttötapausten arvioimiseksi on lisäksi tarpeen määritellä mihin tällainen supertietokone soveltuu ja mitkä sen rajoitukset ovat.

Supertietokone on kymmenien miljoonien eurojen investointi, jonka hankinta, käyttö ja ylläpito vaatii hitaasti hankittavaa asiantuntemusta ja massiivisia ajonaikaisia energiamääriä. Sen käyttöikä on vain muutamia vuosia ennen kuin sen viimeisintä kehitystä edustava teknologia vanhentuu eikä sitä kannata ylläpitää vaan se on korvattava uudella. Jotta supertietokoneinvestoinnista saadaan mahdollisimman suuri hyöty, se pyritään pitämään päällä jatkuvasti ja sen käyttöasteen tavoite on 85–95 %.

Tähän tavoitteeseen voidaan päästä, kun supertietokoneen käytönhallinta-järjestelmässä on jatkuvasti suuri joukko laskentatöitä. Pääosa supertietokoneen toiminnasta tapahtuukin ns. eräajojärjestelmän kautta (Kuva 2), joka automaattisesti valitsee kullekin laskentatyölle tarvittavat laskentanoodit ja niissä tapahtuvat laskentatyöt. Se poimii ne eri tarkoituksiin soveltuvista ns. jonoista, joihin käyttäjät lähettävät laskentatyönsä. Käyttäjien tehtävä on varmistaa, että töiden vaatimat ohjelmat toimivat optimaalisesti ja että niiden vaatimat aineistot on ladattu tallennuslevyille. Supertietokoneessa voidaan myös ajaa interaktiivisia töitä, mutta yleensä hyvin rajatusti.



Kuva 2. ArchiMate-arkkitehtuurikuva tyypillisen HPC supertietokoneen toiminnallisista osista.

Supertietokoneen toiminnot vaativat, että laskennan kohteena olevan **aineiston on oltava sen omassa levytilassa**, josta automaattitoiminnot siirtävät sen tehokkuuden saavuttamiseksi mahdollisimman lähelle suorittimia (lukurajoitusten vähentäminen).

Eräajojärjestelmä aiheuttaa vaikeasti ennustettavan viiveen ennen kuin työt siirtyvät laskentaan. **Yleensä HPC ei sovellu interaktiiviseen käyttöön**. Viiveen ennustettavuutta HPC:ssä pyritään parantamaan määrittelemällä jonoja, joilla on erilainen maksimikapasiteetti ja ajoaika, joista käyttäjät valitsevat sopivimman. Toisaalta **erittäin suuret ja pitkäaikaiset työt vähentävät järjestelmän joustavuutta ja tehokkuutta**, sillä ennen näitä ajoja ja niiden aikana muut työt joutuvat odottamaan hyvinkin pitkiä aikoja. Tällaiset työt vaativat usein erityisjärjestelyjä.

4. Käyttötapaukset ja niiden vaatimukset

Käyttötapaukset

Työpajan osallistujilta saatiin yhteensä 14 kirjallista käyttötapausta, jotka kaikki kuvasivat tähän asti täyttymättömiä laskentatarpeita (Liite B. Käyttötapaukset). Ne vaihtelivat hyvin yleisestä erittäin spesifiin.

Työpajakeskustelussa käyttötapauksen tarpeita verrattiin HPC:n kyvykkyyksiin ja arvioitiin soveltuvuutta.

Yleisimmäksi tarpeeksi nousi grafiikkasuorittimien (engl. Graphics Processing Unit, **GPU**) käyttö. Grafiikkasuorittimia käytetään nykyään laajaan kirjoon

laskentaongelmia, jotka ovat voimakkaasti toistuvia ja rinnakkaistuvia. HPC-ympäristöissä niiden käyttö tarjoaa ylimääräisen rinnakkaistuvuusulottuvuuden. Modernit GPU-piirit ovat myös hyvin kalliita ja niiden käyttö jaetuissa ympäristöissä on usein perusteltua. Toisaalta käyttötapauksen tarkastelussa on hyvä käsitellä pilvipalveluita toisena, HPC:stä poikkeavana suurteholaskennan alueena. Pilvipalvelut voivat myös palvella organisaatioiden GPU-laskentatarpeita.

Eri suorittimia käyttävät hyvin laajat ja rinnakkaistuvat laskentatehtävät ovat HPC:lle ominaisimpia käyttötapauksia. Laajaksi katsomme CSC:n nykyisen supertietokoneen Mahdin töiden keskimääräisen koon, 24 GPU:ta, ylityksen. Erittäin pitkäkestoisin töihin voidaan myös rakentaa vaiheistus, joka käytännössä mahdollistaa niiden laskennan pysäyttämisen ja myöhemmin jatkamisen ('checkpointing'). Tämä tulee tyypillisesti tarpeelliseksi yli vuorokauden jatkuvissa töissä.

Näitä lähestymistapoja käyttäen pystyimme jakamaan esitetyt käyttötapaukset neljään pääluokkaan (Taulukko 2).

Taulukko 2. Sidosryhmien toimittamat käyttötapaukset jaettuna neljään pääluokkaan. Luokat eivät ole toisiaan pois sulkevia.

Luokitus	Kuvaus	Lisäselitys	Käyttötapaukset
1. Laaja	Laajat GPU-resurssitarpeet / HPC-eräajoihin sopivat	Raskas GPU/CPU-laskenta, tehtävät pilkottavissa	1, 9, 10, 12, 13
2. Interaktiivinen	Interaktiivinen käyttö		2, 3, 7, 8, 11
3. Pilvi	Dedikoitu pilvipalvelu / organisaatiokohtaiset GPU-ympäristöt		1, 5, 8, 14
4. Pitkä	Pitkäkestoiset ajot	Tehtävien jakaminen (Checkpointing)	1, 13, 14

Vaatimukset

Käyttötapaukset edellyttävät hyvin erilaisia käyttäjäpolkuja ja osaamista. Osa käyttötapauksista vaatii syvää HPC-osaamista, kuten laskennan optimointia, ajon pilkkomista ja konttipohjaisten ratkaisujen hallintaa. Datan tuottajille laskenta

muodostaa vain lopullisen askeleen pitkällisessä ja työläässä ketjussa aineistojen saatavuuselvittelyjen, keräämisten ja valmistelujen kokonaisuudessa.

Palvelun tuotteistamisessa korostuu siksi tarve selkeille ja yhtenäisille käyttäjäpoluille, kattavalle ohjeistukselle sekä yhteisille toimintamalleille, jotta käyttäjät voivat arvioida tarpeensa käyttää HPC-ympäristöjä mahdollisimman tehokkaasti ja turvallisesti. Luonnollisesti ohjeistus pyrkii myös tekemään mahdollisimman selväksi mihin HPC ei sovellu. Tutkijoiden osaamistaso vaihtelee, eikä HPC ole kaikille tuttu toimintaympäristö, minkä vuoksi tarvitaan synteettistä dataa sekä erillisiä harjoitteluympäristöjä, joissa käyttäjät voivat turvallisesti opetella HPC-ympäristöjen, työkalujen ja työnkulkujen käyttöä ennen varsinaista tuotantokäyttöä.

Oheinen taulukko 3 antaa arvioita terveystietojen tuottajien lähitulevaisuuden tarpeista.

Taulukko 3. Arvio käyttötapauksen toistuvuudesta ja volyymista vuodelle 2027.

Varhan Auria tietopalvelun arvio: 5–10 uutta CSC GPU-ympäristöä; tyypillinen tarve 1–4 x Nvidia H100 (96 GiB) useiksi viikoiksi, kuvantamisessa jopa 8 GPU:ta kuukaudeksi
--

HPC:n laskentaresurssien hyödyntämisen edellytyksenä on niiden kytkeytyminen **laajempaan ekosysteemiin**, jota arkaluonteisen aineiston käsittelyn haasteet monimutkaistavat. Aineisto on ensinnäkin **siirrettävä** HPC-järjestelmän sisään. Tämän täytyy toisilain mukaan tapahtua käyttäen riittävää **suojaustasoa** ja **tietoturvallista käyttöympäristöä**. Jotta siirto olisi mahdollista, tulisi käyttöympäristöjen ja niiden käyttäjien seurata jaettuja sääntöjä. Tarkastelemme niitä seuraavissa luvuissa.

5. Oikeudellinen yhteentoimivuus

Keskeinen sääntely

Tämän raportin toimeksiantona on CSC:n HPC-supertietokoneiden käyttö toisilain alaisuudessa, joka on juuri muutettu vuoden 2019 alkuperäisestä muodostaan³. Alkuperäinen laki loi sosiaali- ja terveystietojen tietolupaviranomaisen Findatan, jolla on laajat valtuudet ja vastuu terveystietojen toisiokäytön luvittamisesta ja valvonnasta. Lisäksi muut alan toimijat ovat voineet luvittaa omaa aineistoaan toisiokäyttöön.

³ Laki sosiaali- ja terveystietojen toissijaisesta käytöstä <https://www.finlex.fi/fi/lainsaadanto/saaduskokoelma/2019/552>
<https://www.finlex.fi/fi/lainsaadanto/saaduskokoelma/2025/767> <https://findata.fi/uutiset/mika-toisilaissa-muuttui-katso-kooste-uusista-ohjeista/>

Muutoksen myötä muut toimijat voivat tiettyjen ehtojen täyttyessä luvittaa omaa aineistoaan myös projektiin, jossa yhdistetään tietoja useasta lähteestä.

Toisiolaki määrittelee toisiokäytön periaatteet, jotka pohjaavat henkilötiedon suojaamiseen ja levittämisen estämiseen tietoturvallisen käyttöympäristön avulla. Findata on antanut määräyksen tietoturvallisen käyttöympäristön vaatimuksista. Suomessa on nyt 10 hyväksyttyä käyttöympäristöä⁴.

Suomen toisiolaki on vaikuttanut vahvasti Euroopan Terveysdata-alueen (European Health Data Space, EHDS) vaatimuksiin, jotka tulevat vaiheittain velvoittaviksi lähivuosina. Terveyslupaviranomaisen ohjaava, koordinoiva ja välittävä rooli Suomen sisällä ja EHDS:n eurooppalaiseen keskusjärjestelmään päin vahvistuu. EHDS-säätely johtaa myös Euroopan laajuisen tietoturvallisen käyttöympäristön (Secure Processing Environment, SPE) käyttöönottoon.

EHDS:n yhteydessä tapahtuvasta SPE:n määrittelystä on jo seurannut, että se nähdään yleispäteväenä ratkaisuna sekä yleisen tietosuoja-asetuksen kuvaamiin korkeampaa tietosuojatasoa vaativien aineistojen että **kaupallisten tietojen käsittelyyn**. Samoin EU:n tekoälyasetuksen (AI Act) velvoittama **tekoälyn sääntelyn testiympäristö** ('AI regulatory sandbox') nähdään samankaltaisena⁵.

Näiden eri käyttötapauksen ristipaineeseen reagoiminen selkeyttävällä ja yhteneväistävällä tavalla on lähitulevaisuuden tärkein tavoite. Tähänastinen säätely on rakennettu yksinkertaisten käyttötapauksen pohjalle. Tutkimusprojektit on nähty tiukasti ajallisesti ja paikallisesti rajattuina prosesseina, joissa rakenteisesta rekisteritiedosta tuotetaan aggregointiin pohjaavia anonyymejä tuloksia. Omiikka-aineistojen kompleksisuus, suuret aineistokoot, maantieteellinen hajautuneisuus, tarve yhdistellä tietoa poliittisten ja organisatoristen raja-aitojen ylitse, sekä tekoälysovellusten räjähdysmäinen kehitys vaativat erilaista lähestymistapaa.

Terveystietojen arkaluonteisuus asettaa sen käsittelylle vielä erityisiä haasteita, joihin täytyy vastata kokonaisvaltaisesti, jotta sen käsittelyn turvallisuus voidaan taata. Käytännössä tämä lähtee tietokonejärjestelmissä niitä käyttävien ihmisten tunnistamisella. Euroopan uudistettuun eIDAS-sääntelyyn⁶ ('electronic Identification, Authentication and trust Services') sisällytetty **digilompakko** tarjoaa merkittävän edistysaskeleen. Korkean tason tunnistautuminen digilompakon avulla pitäisi olla tämän vuoden loppuun mennessä kaikkien eurooppalaisten käytettävissä. Sen jälkeen digilompakon käyttömahdollisuudet laajenevat organisaatioihin ja muuhun käyttöön kuten tutkimukselle tärkeään käyttöoikeuksien hallintaan. Sallivalla lainsäädännöllä on myös mahdollista ottaa käyttöön kolmansien osapuolien digilompakot, joka

⁴ Astori <https://astori.lv.fi/>

⁵ Lähteenmäki, J., & Pajula, J. (2025). *Tekoälyn sääntelyn testiympäristöt sote-innovaatioiden vauhdittajina*. VTT Technical Research Centre of Finland. VTT Tutkimusraportti No. VTT-R-00556-25 <https://cris.vtt.fi/en/publications/teko%C3%A4lyn-s%C3%A4%C3%A4ntelyn-testiymp%C3%A4rist%C3%B6t-sote-innovaatioiden-vauhdittaj/>

⁶ <https://dvv.fi/mika-on-uudistettu-eidas-asetus>

mahdollistaisi tämän teknologian nopeamman ja laajemman hyväksikäytön tutkimuksen mahdollistajana.

Nykyisten tutkimuksen luvitusjärjestelmien liittäminen virallisiin tunnistamisjärjestelmiin on välttämätöntä, sillä HPC on edellä kuvatulla tavalla lähtökohtaisesti itsenäinen prosessointiympäristö, jonka käyttäminen aineistoa tiukasti suojaavalla tavalla vaatii tiukasti yhteen nivotun tietoteknisen ekosysteemin, josta yleisesti käytetään nimitystä federaatio tai uudemmalla EU-terminologialla tietoalue ('data space').

EHDS-sääntely ei nimestään huolimatta anna ohjeita EU:n jäsenmaiden sisäiseen tai väliseen lupien ja tietojenvaihdon koordinaatioon sen paremmin kuin Suomenkaan toisiolaki. Parhaillaan tutkimuskäytössä olevat eurooppalaiset biotieteen verkoston ELIXIR:in LS Login ja GÉANT:in MyAccessID joutuvat sopeutumaan digilompakkokauteen, jossa lupien välitys pohjaa uuteen tekniseen ratkaisuun ('verifiable credentials'⁷). Tätä hahmotellaan laajan EU-vetoisen konsortion Data Spaces Support Centre'n uusimmassa arkkitehtuurissa⁸, joka ei valitettavasti vielä suoraan käsittele arkaluonteista tiedon hallintaa.

Lailliset reunaehdot CSC:n HPC-ympäristöissä

CSC toimii omistajiensa sidosyksikkönä EU:n julkisia hankintoja koskevan direktiivin (2014/24/EU, 12 artiklan 1 ja 3 kohta) mukaisesti ja tuottaa palveluja ensisijaisesti omistajilleen. Vaikka eduskunta hyväksyi 4.6.2026 CSC-lain⁹, joka lisää joustavuutta erityisesti suurteholaskennan resurssien ja infrastruktuurin hyödyntämiseen, CSC toimii edelleen hankintalainsäädännön puitteissa: myynti ulkopuolisille on rajattua ja kohdistuu tutkimus-, kehitys- ja innovaatiotoimintaan (TKI).

Kansallisten supertietokoneiden lisäksi CSC ylläpitää myös **EuroHPC**-yhteisyrityksen supertietokoneita. Näissä on merkittävä yhteinen EU- ja yksityinen rahoitus ('European High-Performance Computing Joint Undertaking', EuroHPC JU), joista laskentakapasiteettia on myytävissä vähäisemmin rajoituksin.

Näistä ensimmäinen on LUMI supertietokone, johon tutkijoilla on kaksi väylää hakea EuroHPC-yhteisyrityksen ylläpitämiä LUMI:n resursseja: kaikki eurooppalaiset tutkijat voivat hakea niitä EuroHPC:n kiintiöstä, johon on allokoitu puolet kapasiteetista ja konsortion jäsenmaiden tutkijat voivat hakea resursseja myös oman maansa kiintiöstä. Tutkijat, jotka ovat affilioituneet akateemiseen tutkimusorganisaatioon tai yrityksiin, jotka on perustettu tai sijaitsevat LUMI-konsortiomaassa, voivat hakea maakohtaisia resursseja. Valitettavasti LUMI-supertietokoneen rakenteet estävät sen auditoinnin

⁷ Verifiable Credentials Data Model v2.0 <https://www.w3.org/TR/vc-data-model-2.0/>

⁸ DSSC Blueprint v3.0 <https://toolbox.dssc.eu/>

⁹ CSC-lain eduskuntakäsittely <https://www.eduskunta.fi/asiat-ja-aanestykset/valtiopaivaasiat/HE%2018%2F2026%20vp>

toisioaineistojen käyttöön. Tulevan LUMI AI -supertietokoneen hankinnassa pyritään täyttämään toisiolain vaatimukset.

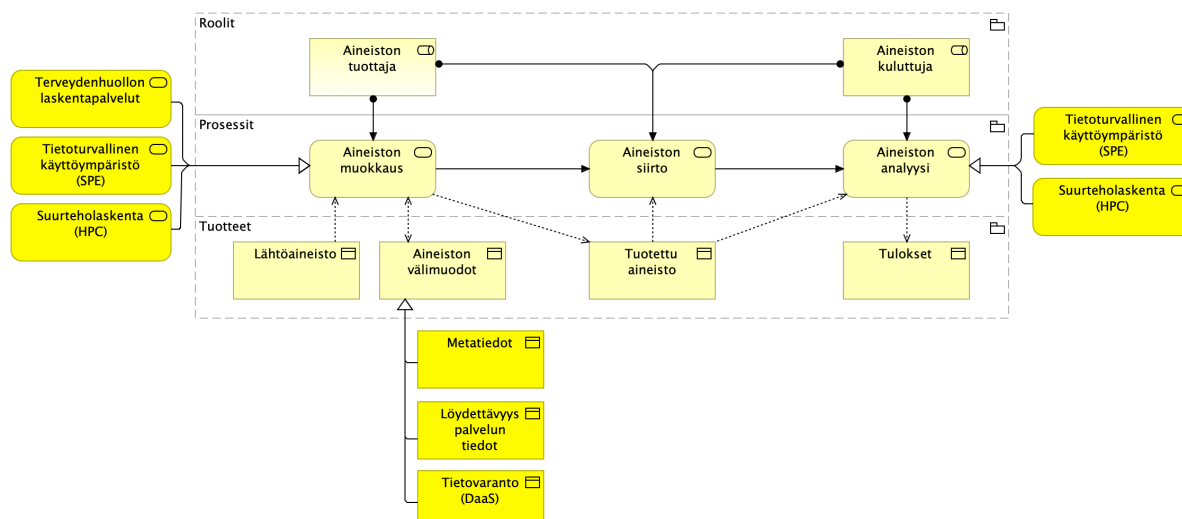
HPC:n toisiokäytön kannalta merkittävin kansallinen ero on seurausta eurooppalaisesti poikkeuksellisen tiukasta tutkimuksen lakipohjaisesta jaottelusta lääketieteelliseen ja akateemiseen. Kun akateemisesta tutkimuksesta vastuullinen Opetusministeriö rahoittaa CSC:n supertietokoneresursseja, niin Sosiaali- ja Terveysministeriön alaisilla hyvinvointialueilla ja niiden tutkijoilla ei ole oikeutta käyttää kansallisia supertietokoneita. Tämä historiallinen jako on niin syvällä rakenteissa, että sitä ei ole vielä voitu ratkaista, vaikka kansallisten resurssien järkevä jakaminen ja tutkimusmenetelmien yhtenäistyminen digitalisaation myötä sitä puoltaisikin.

Findatan määräyksen kohta 2.5.1.2. on yleisesti tulkittu tarkoittavan tietoturvalisille käyttöympäristöille jatkuvaa 24/7 valvontaa ja reagoitua. Se nostaa niiden ylläpitokustannuksia moninkertaiseksi tuottamatta merkittäviä hyötyjä, sillä normaalillakin seurannalla ongelmatilanteet saadaan haltuun nopeasti ja lopulliset ratkaisut vaativat usean tahon yhteistoimintaa, joka on mahdotonta normaalin työajan ulkopuolella. Tämän vaatimuksen laajentaminen jaettuun tutkimuskäytön HPC-ympäristöön olisi suhteettoman kallista. Tästä vaatimuksesta tulisi luopua.

6. Organisatorinen yhteentoimivuus

Käytännön tasolla hyvinvointialueiden tutkijat muun tutkimusyhteisön palveluista erottava tekijä on elektroninen tunniste, joka ei kuulu korkeakoulujen ja tutkimuslaitosten Haka-järjestelmän¹⁰ piiriin. Hyvinvointialueiden toimijoista yliopistosairaalat ovat erikoisasemassa, sillä monella niiden tutkijalla on myös yliopistotunnus. Haka-tunnuksella varustetulla yliopistosairaalan tutkijalla on näin jo mahdollisuus käyttää CSC:n HPC-palveluita tarpeen mukaan lakien sallimissa rajoissa. Toisiolain piirissä HPC-palveluiden käytön tulee siis ottaa huomioon kaksi pääkäyttötapausta, joista aineistojen tuottajien käyttö on rakenteellinen haaste (Kuva 3). Tutkimuskäyttöön kehitettyä arkaluonteisen HPC-käytön palvelun (SD HPC) semanttisia ja teknisiä ratkaisuja kuvataan myöhemmissä luvuissa. Tässä käsitellään aineistojen tuottajien näkökulmaa HPC-palveluihin ja niiden käytettävyyteen.

¹⁰ Haka <https://wiki.eduuni.fi/spaces/CSCHAKA/pages/27297787/Luottamusverkosto>



Kuva 3. ArchiMate-arkkitehtuurikuva HPC:n käytön kahdesta erillisestä käytöstä toisilain alaisuudessa. Rakenne seuraa tietolueiden yleisarkkitehtuuria korostaen laskentapalveluja ja arkaluonteisten tietoaineistojen välivaiheita.

Hyvinvointialueet

LUMI:lla on sekä kansallinen että EuroHPC JU-kiintiö¹¹. Kansallisesta 20 % on varattu kaupallisten toimijoiden TKI-toimintaan. Toistaiseksi LUMI:lla ei ole vielä ollut yhtään hyvinvointialuetta asiakkaana, koska haasteena on LUMI:n soveltumattomuus sensitiivisen datan käsittelyyn tällä hetkellä. Tietenkin dataa voidaan anonymisoida ennen LUMI:lle siirtoa, mutta se ei ratkaise ongelmaa, sillä anonymisointikin on sen tyyppistä raskasta laskentaa, johon HPC:stä toivotaan apua.

Hyvinvointialueilla on kolme vaihtoehtoa tulla LUMI:n käyttäjäksi, kaksi maksuttoman käytön ja yksi kaupallisen käytön tapaus.

Maksuton käyttö on mahdollista seuraavissa tapauksissa:

- Hyvinvointialue tekee yhteistyötä suomalaisen yliopiston tai tutkimuslaitoksen kanssa. Käyttö on maksutonta silloin, kun PI tulee yliopiston tai tutkimuslaitoksen puolelta ja tulokset julkaistaan tieteenalan käytännön mukaisesti.
- Hyvinvointialue hakee EuroHPC JU:n hakujen kautta kapasiteettia tutkimukseensa. Mikäli heille myönnetään kapasiteetti, niin käyttö on maksutonta, mutta käytöstä ja tuloksista pitää raportoida EuroHPC JU:lle.

Lisäksi hyvinvointialueet voivat tehdä kaupallisen sopimuksen, jolloin käyttö menisi siihen 20 % kiintiöön, jota kaupalliset toimijat voivat käyttää TKI-työhönsä. Tällöin hyvinvointialueen kanssa tehdään sopimus, jossa määritellään kuinka paljon kapasiteettia he haluavat ostaa ja heiltä laskutetaan kapasiteetista käytön mukaan listahinta¹². Kaupallisessa käytössä ei ole julkaisuvaatimuksia, vaan hyvinvointialue voi itse päättää mitä tuloksilla tekee.

¹¹ LUMI General Terms of Use https://lumi-supercomputer.eu/wp-content/uploads/2026/03/LUMI-General-Terms-of-Use_2026.pdf

¹² LUMI:n hinnasto <https://csc.fi/tietoa-meista/asiakkaat/suurteholaskentaa-yrityksille/#palveluhinnasto>

Kansallisesti koordinoitu toimintamalli

Suomen kansallisen terveystietoalueen ehdottama yhtenäinen datainfrastruktuuri toteutettuna avoimilla standardeilla ja ohjelmistoilla vastaisivat hyvin laajemman HPC-käytön vaatimuksia. Samoin toisiolain uudistuksessa hyvinvointialueiden ja muiden kansallisten rekisterinpitäjien suuremmat toiminta- ja päätöksentekovaltuudet vaativat tiukempaa koordinaatiota toimijoiden välillä.

Toisiolaki sekä Findatan rooli lain käytäntöjen koordinoijana rakentaa jo pitkälle Euroopan data-alueiden yleisarkkitehtuurin mukaista säännöstöä ('Governance policy' ja 'Rulebook', Kuva 4). Toisiolaki ja tulevan Euroopan Terveystietoalueen (EHDS:n) määrittelevät roolit pyritään nykysuunnitelmien mukaan pitämään Suomessa pitkälti yhdenmukaisina.

Selkeä lakipohja myös helpottaa rakentamaan kansallisen verkoston vaatimaa korkeaa luottamustasoa. Jotta se kääntyy myös käytännön tasolle, täytyy luottamusta tukevat toimenpiteet määritellä tarkemmin kuin tällä hetkellä.

Toiminnallinen tietoalue (federaatio) koostuu eri rooleja toteuttavista jäsenistä. Luonnollisten henkilöiden tunnistamisen lisäksi myös nämä organisaatioroolit tulee voida yksiselitteisesti ja luotettavasti tunnistaa. Henkilötietoa käsittelevän tietoalueen yhteisten hallintopalveluiden tulee kattaa sen jäsenten (tiedontuottajien, palvelutuottajien, tiedonkäyttäjien) lisäksi myös aineistot, luvat ja niiden mukaan määrittyvät projektit, sekä palvelut. **Tietoalueen tulee ylläpitää näistä rekistereitä**, joista jäsenten ylläpitämät palvelut voivat varmistaa tietoalueen toimintojen oikeudellisuuden.

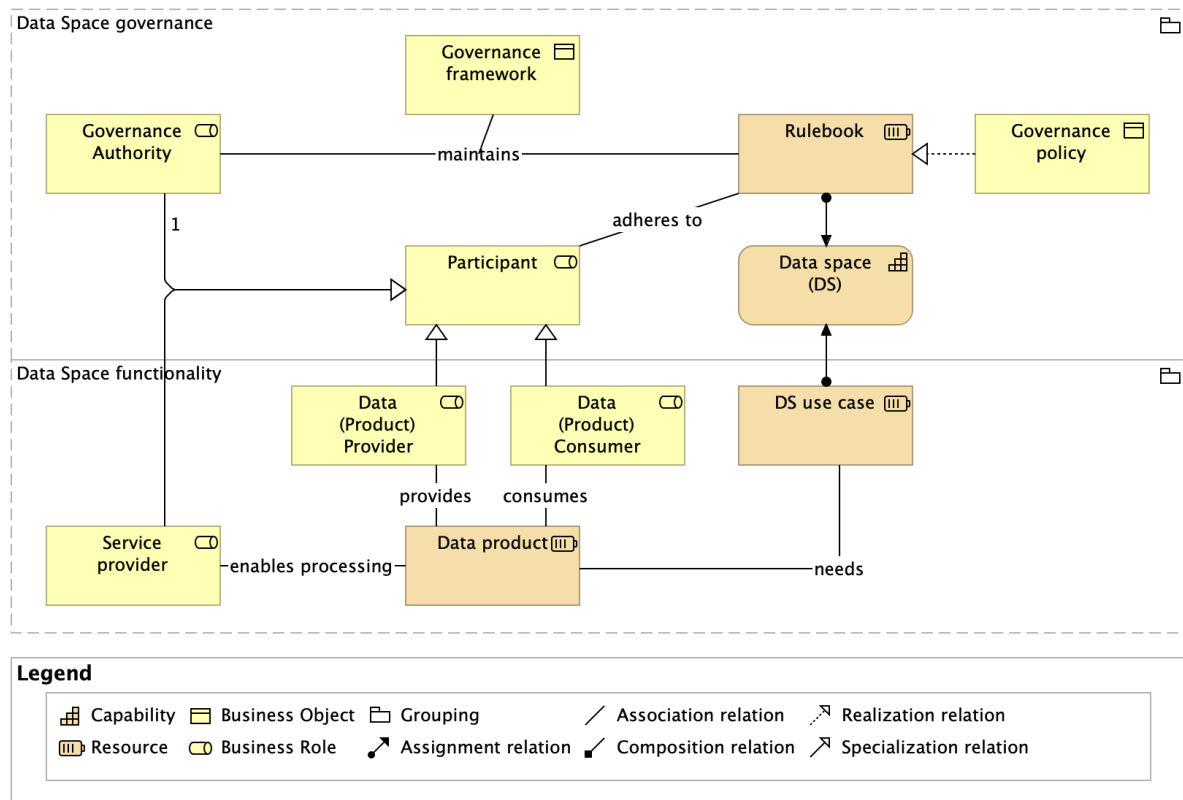
Näiden lisäksi **henkilötiedon siirron turvatasosta** palveluihin ja palveluiden välillä tulisi sopia. Turvatason mukaisia kaikkien toimijoiden käytössä olevia palveluita ja protokollia pitäisi olla vähintään yksi. Tällä hetkellä protokollasta ei ole ohjeistusta sen paremmin toisiolain alaisessa toiminnassa kuin EHDS:ssäkään. Tämä palvelisi laajemminkin hajautetun laskennan tarpeita. Hajautetun laskennan huomioon ottavan ohjauksen odottaminen on jo johtanut yksittäisten toimijoiden erillisratkaisuihin¹³.

Toisiolakiin on kirjattu velvoite tietoturvallisesta käyttöpalvelusta, joka kuvaukseltaan vastaa Suomi.fi palveluväylän toiminnallisuutta. Lakiin hallinnon yhteisistä sähköisen asioinnin tukipalveluista vuoden 2026 alusta voimaan astunut muutos¹⁴ poisti velvoitteen ja mahdollistaa muiden korkean tieturvatasojen ratkaisujen käyttöönoton.

¹³ HUS Viranhaltijapäätös § 39 / 2024. Hajautettuun analytiikkaan perustuviin hankkeisiin liittyminen ja niiden toteutus HUSissa

¹⁴ Palveluväylän viranomaiskäytön velvoitteen poisto <https://kehittajille.suomi.fi/ajankohtaista/tiedote-kayttovelvoitteen-poistumisesta>

Yleiskäyttöiselle ja turvalliselle tiedonvälitysväylälle on kansallinen selkeä tarve.



Kuva 4. ArchiMate-kaavakuva DSSC Blueprint v.3.0 tietalueiden hallintorakenteesta ja perustoiminnoista. Mukailtu lähteestä: <https://blueprint.dssc.eu/?pane=intro>

Terveystietoaaineistojen luvallinen toisiokäyttö ei tulisi olla vain ihmisluettavan tietoluvan ('data permit') varassa, vaan siitä on muodostettava koneluettava pääsylupa('access permit', 'access token', 'verifiable credential'), jonka avulla voidaan automatisoida luvansaaneiden pääsy luvitettuun aineistoon.

Tiedon tuottajat tarvitsevat aineistojen valmistamiseen yhtenäisen mallin metatietojen kuvaukseen, pseudonymisaatioon ja anonymisointiin. Jälkimmäisin toiminto on myös tarpeen terveystiedon käyttäjälle hänen ottaessaan tuloksia ulos tietoturvalisesta käyttöympäristöstä.

Tietoalueen sisällä pääroolien käyttämät palvelut muodostavat osa-alueita, joiden tulee pystyä kontrolloidusti kommunikoimaan toistensa kanssa. Toisiokäyttäjien SPE:iden tulee pystyä tietoluvallisesti jakamaan tehtäviä muiden käyttöympäristöjen kanssa muodostaen projektikohtaisia federaatioita ('federated computing') tai käyttää muita palveluita kuten HPC:itä. Tiedon tuottajien prosessointiympäristöjen tulee pystyä sekä turvallisesti kommunikoimaan toistensa kanssa heitä koskevista tietolupahakemuksista, että helposti jakaa keskenään valmistelemiaan osa-aineistoja, jotka koordinoiva luvittaja yhdistää luovutettavaksi kokonaisuudeksi ('data linkage'). Lopputulokseen pääsemiseksi tiedontuottajien tulee käyttää henkilö- ja muiden tunnisteiden muuttamiseksi yhteistä pseudonymimuunnosta, joka tulee kirjata **lupakohtaiseen pseudonymiavainrekisteriin**, josta se myös parhaassa tapauksessa jaetaan aineistoa valmistelevien kesken.

Suurteholaskennan kannalta olisi hyödyllistä, että kun tiedontuottajat työstävät terveystiedosta erilaisia välivaiheita, ne olisivat tai päätyisivät nopean tietoliikenneyhteyden päähän HPC:stä ympäristöön, josta edellä mainittuja työstettyjä välivaiheita olisi myös mahdollista jakaa.

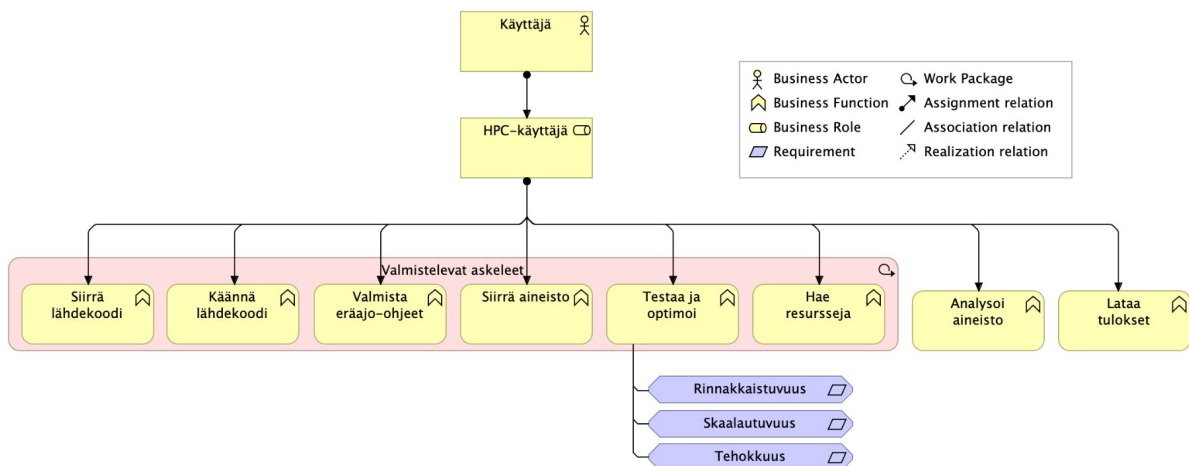
Tämä visio pohjaa olemassa olevaan CSC:n SD Desktop ja SD Connect -palveluiden avoimiin standardeihin perustuviin toiminnallisuuksiin. SD Desktop on SPE:n interaktiivinen laskentakomponentti ('virtual desktop'), joka kytkeytyy HPC:hen ja SD Connect on aineistojen tallennusosa ('object storage'). Nykyinen SD Connect, kun sitä käytetään tutkijakäyttäjien itsensä hallinnoiman tiedon tallentamiseen, sisältää mahdollisuuden jakaa kansioita tiedostoineen toisille projekteille. Tiedon tuottajien versio SD Connectista voisi jakaa tietoluvan mukaisia osa-aineistoja hakemusta koordinoivalle taholle. Tätä tukee myös kuvassa 3 mainittu tietovaranto-konsepti ('Dataset-As-A-Service', DaaS), jonka mukaan levitykseen tarkoitettut aineistot työstetään kokonaisuudessaan muotoon, josta on ohjelmallisesti mahdollista ottaa luvitettavia ja luovutettavia osa-aineistoja.

Palveluprosessit ja tulevaisuus

Palveluiden avoimet standardit tarkoittavat, että pitkällä aikavälillä näiden palveluiden ei tarvitse toimia pelkästään CSC:n käsittely-ympäristöissä. Kun yhteensopivuus saadaan kattavammaksi, on useampien tiedontuottajan täysin mahdollista luvittaa aineistoja eri SPE:hin prosessoitavaksi. Samoin aineistoja voidaan luvittaa eri alustoilta.

Alusta ja datankäsittely-ympäristö eriytyvät myös seuraavan sukupolven SPE:n myötä. Trendi Euroopassa on nyt laskentaympäristön siirtyminen kontteihin ('container'), mihin suuntaan SD Desktop-palveluakin kehitetään. Tavoitteena on modulaariset konttiympäristöt, joissa yhteen liitetään sekä aineistoja että ohjelmia sisältäviä kontteja ja tämä kokonaisuus voidaan lähettää laskettavaksi halutulle alustalle. Tarvitaan yhteisesti sovittu ja hyväksytty prosessi uusien konttien ja muiden työkalujen tietoturvan tarkastamiseen, jotta ne voidaan ottaa käyttöön toisiolain alaisia aineistoja käsittelevissä ympäristöissä. Samoin olisi hyvä luoda yhteinen rekisteri, jossa olisi tarjolla jo hyväksyttyjä valmiita kontteja, joista valita.

Laskentaympäristöjen monipuolistuminen antaa mahdollisuuksia niille erikoistua tarjoamaan erikoistuneita palveluita. Näistä HPC on vain yksi, jonka haasteet sen käyttäjille poikkeavat muista ympäristöistä (Kuva 5). Perinteisesti käyttäjä on ollut vastuussa lähdekoodin optimoinnista jokaiseen uuteen ympäristöön. Myöskään ei ole ollut itsestään selvää, että kaikki projektit saavat käyttöluvan HPC-ympäristöön, mikäli koodia ei ole optimoitu sekä tarpeeksi rinnakkaistuvaksi ja skaalautuvaksi. Arkaluonteiset aineistonkäsittelyt tuleekin testata synteettisillä aineistoilla. Käyttäjät, sekä yksilöt että organisaatiot, tulevat tarvitsemaan sekä **koulutusta että tukipalveluita** HPC:n käyttöön.



Kuva 5. Yleinen HPC-käyttäjän polku, joka korostaa valmistelevan työn merkitystä HPC:n käytössä.

7. Semanttinen yhteentoimivuus

Tieturvallinen käyttöympäristö on lähtökohtaisesti palvelu, joka antaa käyttäjille välittömän pääsyn arkaluonteiseen aineistoon. Käyttäjät kuuluvat projektiin, joka säätelee mihin palveluihin ja aineistoihin käyttäjillä on pääsy. Toisistaan eristetyt projektiympäristöt koostuvat laskenta- ja tallennusosiosta.

Jaettua HPC-ympäristöä ei voida kokonaisuutena liittää projektin käyttöympäristöön, vaan HPC:stä täytyy voida hetkellisesti eristää osa projektin käyttöön tiedon suojaamiseksi. Mikäli sama palveluntuottaja tarjoaa molemmat palvelut, on liittäminen jaetun sisäverkon, projektinhallinnan ja aineistonsiirron kannalta helpompaa kuin jos HPC olisi täysin erillinen. Molemmissa tapauksissa arkaluonteisen tiedon suojaamisen vaatimukset ovat niin merkittävät, että HPC-käyttö voidaan nähdä esiasteena hajautetusta ja palveluntarjonnasta. Epäsuoraksi ('indirect query') palvelun tekee se, että käyttäjän arkaluonteisen HPC:n palveluun lähettävä eräajopyyntö ei sisällä käsiteltävää aineistoa suoraan, vaan vain viitteen sen paikantavaan tunnisteeseen, jota käyttäen aineisto on haettava erikseen.

Toisiokäytössä aineiston käyttöoikeus on määritelty tietoluvalla. Luvitettua aineistoa pääsee käyttämään vain tietoturvalisessä käyttöympäristössä. Tietoturvalisessä HPC:ssä tapahtuva käsittely ei ole tietoluvan vastaista toimintaa eikä poikkea kuin nopeudellaan ja skaalautuvuudellaan tavallisen virtuaalikonepohjaisen tietoturvalisessä käyttöympäristön sisällä tapahtuvasta käytöstä. Tietoturvallinen HPC-käyttö siis laajentaa käyttöympäristön käsitettä. Tästä johdettuna hajautettu käyttö laajentaa sen koskemaan koko yhteistoimintaverkostoa. Tärkeintä onkin, että verkoston kaikki osat toimivat yhteen samalla turvatasolla ja käyttö tapahtuu vain palveluissa, jotka on mainittu luvassa.

CSC:n SD Desktop ja SD Connect palveluiden muodostama tietoturvallinen käyttöympäristö on rakennettu toimimaan yhdessä CSC:n Roihu HPC:n tietoturallisen osan kanssa. Mikäli muut suomalaiset toisiolainalaiset ympäristöt haluaisivat toimia samoin, tulisi niiden implementoida samat tietoturalliset rajapinnat. Lisäksi yhtenäisen tietoalueen (*data space*) yleiset tukipalvelut, jotka varmistavat tietoturvan ja seurattavuuden tulisi olla käytössä. Näistä ensimmäinen on kaikkien käyttäjien luotettava tunnistaminen.

Suurteholaskennan oleellinen osa on suuret aineistot. Niiden siirtäminen HPC:n sisäiseen tallennukseen on välttämätöntä. HPC-analyysin eniten aikaa vievä osa riippuu aineiston alkuperäisen sijaintipaikan saavutettavuudesta. On siksi **suositeltavaa, että toisiokäyttöön tarkoitetut aineistot sijaitsisivat mahdollisimman lähellä eli nopean siirtoyhteyden takana**. Niiden on myös seurattava yhteneväistä suojausmenetelmää.

Terveystiedon toisiokäytön ominaispiirre on lähtömateriaalina käytetyn terveystiedon vaihtelevuus ja hajanaisuus. Terveystiedon potilastieto on harvoin suoraan toisiokäyttöön sopivaa ilman perusteellista puhdistusta ja harmonisointia. Tätä hidasta ja työlästä vaihetta, joka on useimmiten datan tuottajan vastuulla, tulisi koordinoitusti rahoittaa. **Tietovaranto**-konsepti (Dataset as a Service, DaaS) on kehitetty välivaiheeksi, joka pystyy ylläpitämään karttuvaa tietoa-aineistoa ja käyttämään kyselyrajapintoja loppukäyttäjälle luovutettavien osa-aineistojen poimimiseen. Eri terveystietojen aineistontuottajien tietovarannot olisi hyvä sijoittaa lähellä toisiaan ja olla osa laajempaa kokonaisuutta, jossa ne on helpompi liittää yhteen muodostamaan yhtenäisen kokonaisuuden, joka luovutetaan toisiokäyttäjälle. Nykyinen toisiolaki ei tunnista tällaista käyttötapaa.

Tiedontuottajien yhteentoimiva verkosto tulisi rakentua tietoalueiden yleisten suositusten mukaisesti¹⁵ seuraten **kansallisia ja kansainvälisiä standardeja**. Monet tutkimuksen palvelutoiminnassa käytetyt standardit ovat yhteistyöelin Global Alliance for Genomics and Health (GA4GH) kontekstissa¹⁶ kehitetty. Suomen julkisilla organisaatioilla on mahdollisuus käyttää tiedonsiirtoon kansallista Palveluväylää¹⁷. Findatan Supertunneli-sovellus on tällä hetkellä rakennettu vain tiedontuottajilta Findatalle ja takaisin tapahtuvaan tiedonsiirtoon.

8. Tekninen yhteentoimivuus

Toisiokäytön turvavaatimusten toteuttaminen Roihussa

Roihu-supertietokone on pääkohteena arkaluonteiselle toisiokäytölle. Vaikka näitä tarpeita korostettiin sitä hankittaessa, tarjolla ei ollut sellaista, joka suoraan tukisi sekä normaalia HPC-käyttöä että korkeamman suojaustason toisiokäyttöä. Suojattavan

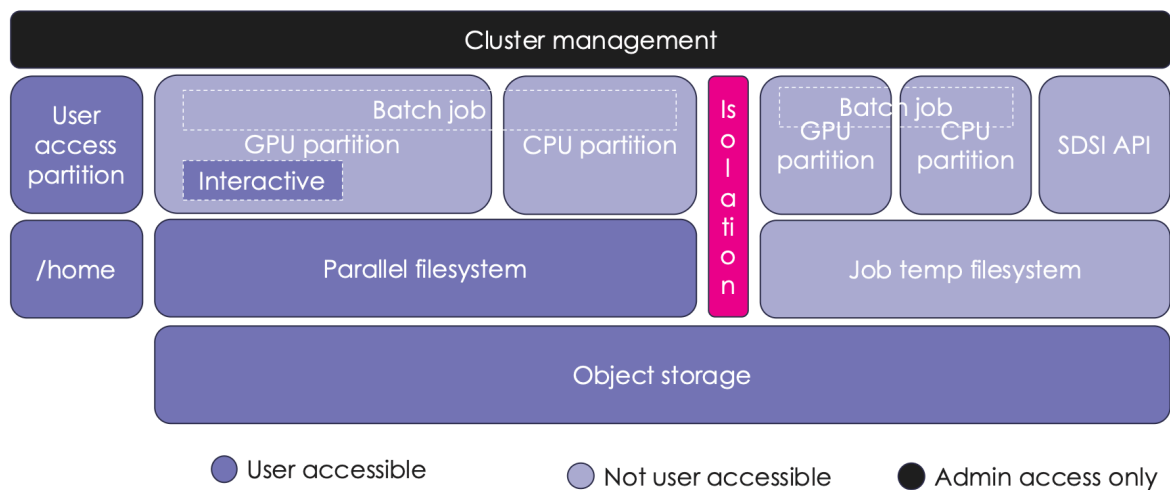
¹⁵ Data Spaces Support Centre Blueprint v.3.0 <https://toolbox.dssc.eu/>

¹⁶ GA4GH standardit <https://www.ga4gh.org/our-products/>

¹⁷ Palveluväylä <https://dvv.fi/palveluvayla>

tiedon käsittely yleiseen käyttöön tarkoitetulla supertietokoneilla on uusi käyttötapaus, johon laitteiston valmistajilla ei ole vielä valmiita ratkaisuja, sillä tarve on vasta noussut esille¹⁸.

Roihun käyttö molempiin tarkoituksiin tullaan mahdollistamaan jakamalla sen resurssit kahteen tiukasti toisistaan eristettyyn osaan (Kuva 6). Julkisesta osasta poiketen käyttäjillä ei tule olemaan suoraan käyttöyhteyttä korkean suojatason laskentaan tai talletukseen vaan kaikki käyttö välittyy SD Desktopista etäohjauksella paikalliseen SD HPC-kontrolliohjelmaan. Alustavien suunnitelmien mukaan korkea suojatase tulee myös johtamaan vaatimukseen kontitetuista sovelluksista ja erillisestä paikallisesta tallennusjärjestelmästä.



Kuva 6. Kaavakuva Roihu-supertietokoneen jakamisesta kahteen toisistaan täysin eristettyyn, eri turvatason osioon. Käyttäjien interaktiivinen käyttö ('User access partition') on korvattu ohjelmointirajapinnalla ('SDSI API').

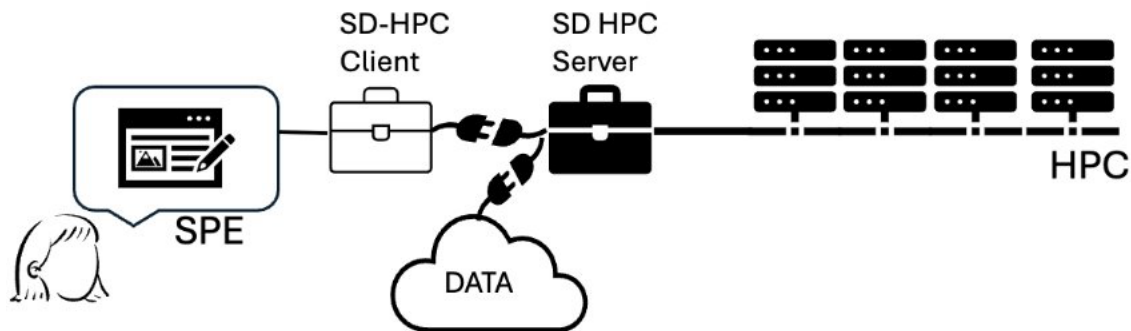
Palveluintegraatio

Tekninen yhteys ja prosessi, SD HPC, jolla sensitiivisiä aineistoja voidaan siirtää tietoturvallisesti HPC-ympäristöön, on testattu nyt väistävällä Puhti supertietokoneella. Se tullaan asentamaan edellä kuvatuin varauksin Roihu ja LUMI supertietokoneisiin.

SD Desktop, käyttäjän interaktiivinen SPE-ympäristö, sisältää ohjelman, HPC Clientin, joka välittää käyttäjän kirjoittaman työnkuvauksen SD HPC Serverille, joka tulkitsee sen, siirtää aineiston HPC:hen ja orkestroi normaalin HPC-eräajojärjestelmän mukaisen ajon ja palauttaa tulokset käyttäjälle (Kuva 7). Oleellista tässä on, että käyttäjä ei

¹⁸ CASC position paper <https://casc.org/policy-publications/position-paper/nist-sp-800-171-guidance-for-research-computing-and-data-centers/>

koskaan pääse jaettuun HPC-ympäristöön vaarantamaan siellä laskettavien aineistojen turvallisuutta.



Kuva 7. Yleinen palveluintegraatio.

Arkaluonteisen tiedon HPC-arkkitehtuuri

Järjestelmän tekninen toteutus perustuu SPE:hen rakennettuun turvallisista rajapinnoista koottuun toiminnallisuuteen, joka välittää HPC-kommunikaatiota molempiin suuntiin (Kuva 8). Tämä tunnetaan Security Serverinä ja voidaan laajentaa mahdollistamaan laajempaa federoitua kommunikointia. Sen vastinpari on rakennettu supertietokoneeseen osaksi SD HPC-serveriä, jonka toteutus tunnetaan nimellä 'Sensitive Data Submission Interface (SDSI) Server'. Tiedostojen haku ja tulosten kirjoittaminen tapahtuu suoraan käyttäjän projektiympäristöön ilman eri yhteydenottoa käyttäjään. Tätä haavoittuvuutta kompensoi krypta-avainten hallintajärjestelmä, joka tuottaa tehtäväkohtaisen avainparin aineistolle. Se vahvistaa entisestään GA4GH crypt4gh kryptajärjestelmän ominaisuutta, jossa kryptattu suurikokoinen tieto on talletettu objektina ja sen purkuun tarvittava otsikko-osa erikseen. Kaikki osat siirretään verkon yli erikseen.

SDSI Serveri orkestroi kaiken tarvittavan sensitiivisen aineiston prosessointiin, ja tämä poikkeaa tavallisen HPC-ympäristön toiminnoista. Se tulkitsee käyttäjän työkuvausten, siirtää sen mainitsevat tiedostot supertietokoneeseen ja lähettää muokatun työkuvausten työjonoon. Kun työ siirtyy ajoon, se siirtää yhä kryptatut tiedostot laskentanoodin talletustilaan ja käynnistää suojauksen purkamisen noodin täydellisen eristämisen jälkeen. Valmiit tulokset kryptataan käyttäjän avaimella ja kun ajo on valmistunut, SDSI Server siirtää tulokset SD Connectiin ja varmistaa, että kaikki ajoon liittyvät tiedostot tuhotaan ennen kuin noodi palautetaan käyttöön.

Alustavasti näemme HPC-palveluiden laajemman toisiokäytön seuraavien vaiheiden kehityksen niin, että ensin mahdollistetaan yhteyden luominen CSC:n ulkopuolisen auditoidun SPE:n ja CSC:n auditoidun SD Roihun välille sekä palveluiden yhteensopivuuden tarkemmalla määrittelyllä. Nämä kehitys- ja käyttökulut tulevat tilaajalle. Tämän jälkeen samanlaista kehitystyötä voisi tehdä myös auditoidun SPE:n ja auditoidun SD LUMI AI:n välille. Tässäkin tapauksessa kehitys- ja käyttökulut tulisivat tilaajalle.

Tekoälytehdas tarjoaa kattavan palveluvalikoiman, johon kuuluvat tekoälyyn optimoidut suurteholaskentaresurssit, data, koulutus ja tekninen tuki. Palvelut ovat maksuttomia tutkijoille, startupeille ja pk-yrityksille. Suuryrityksille palvelut tarjotaan maksullisina.

10. Tiekartta ja suositukset

Tämän projektin aikana työpajoissa on nostettu usein esiin ratkaisujen vaiheistaminen ja selkeä priorisointi kansallisten käyttötapauksien osalta. Korkeimman prioriteetin hankkeena esitetään arkaluontoisen tutkimustiedon, ml. toisiolain alaisten aineistojen, suurteholaskenta CSC:n HPC järjestelmissä akateemisissa projekteissa.

CSC:n palveluiden tiekartalla suurteholaskennan mahdollistaminen SD palveluiden kautta arkaluontoisen aineiston laskentaan arvioidaan alkavan vuoden 2027 alkupuolella. Tätä voidaan pilotoida aineistohaltijoiden kanssa. Toisiolain alaisten aineistojen hallinta vaatii vielä SD palveluiden auditointia käyttötärpeen mukaiseksi. Auditointiprosessin oletetaan tapahtuvan vuoden 2027 aikana.

CSC:n SD palvelut on tarkoitus myös integroida tulevaan LUMI-AI laitteeseen, jonka rakentaminen ajoittuu vuoden 2027 ajalle. Toisiolainkäyttöön auditoitu palvelutoiminta SD palveluiden ja LUMI-AI:n integraationa alkaisi vuoden 2028 aikana. Kappaleessa 5 on kuvattu millaisin ehdoin CSC:n kansallisia sekä EuroHPC:n LUMI palveluita voidaan käyttää maksuttomasti tai miten palveluita voidaan myös ostaa toisiolainalaisten tutkimuksen mahdollistamiseksi. Näissä käyttötapauksissa, esim. hyvinvointialueen, loppukäyttäjä siirtyy käyttämään CSC:n tai EuroHPC:n palveluita, kuten kansallisen akateemisen tutkimusorganisaation tutkijakin toimisi.

Kansallisessa tutkimus- ja innovaatiotoiminnassa tulee huomioida myös muiden kansallisten organisaatioiden tarpeet integroitua CSC:n HPC laskentaan. Projektin työpajoissa nousi esiin tarve miettiä toisiokäyttöympäristöiksi rekisteröityjen palveluiden mahdolliset käyttötapaukset sekä tunnistaa näiden potentiaaliset käyttäjämäärät. Kaikkien käyttöympäristöjen ei ole tarve ylläpitää integraatiota suurteholaskentaan taustalla. Tämän sijaan suositeltiin palveluiden kohdentamista käyttötapauksien mukaiseen toimintaan oikean kohderyhmän tarpeiden täyttämiseksi.

Mikäli yhteisen palvelutoiminnan vaatimukset täyttyvät, yhteinen tiekartta muodostuu palveluvastuiden, yhteentoimivuuden sekä kustannusten tarkemmasta suunnittelusta. Kyse ei siis ole yksinkertaisesta teknisestä API rajapinnasta, joka määrittää organisaatioiden välillä suurteholaskennan mahdollistamiseksi olemassa olevan käyttöympäristön taustalle. Tämän lisäksi kansallisen tiekartan osaksi listattiin tarve yhteisille tietoturvalisille ratkaisuille, kuten tiedon siirrot organisaatioiden välillä, luottamuspalvelut käyttäjien tunnistamiseksi sekä koneluettavan käyttöluvan tuottaminen tietoluvasta. **Kansallinen ekosysteemi** vaatii myös lisäarvopalveluita, kuten aineistojen pseudonyymiavainpalvelu tai salausavainten hallinta. Kansalliselle tiekartalle ehdotetut asiat ovat yhteensopivia FHDS ja EHDS vaatimusten kanssa. Asioita tulisi edistää askel kerrallaan ja parantamalla palvelutoiminnan edellytyksiä iteratiivisesti.

HPC on malliesimerkki kansallisista resursseista, joiden kehittämisessä pitää ottaa huomioon laajasti eri hallinonalojen tarpeet, jotta niiden hyödyt jakaantuvat järkevästi.

Liitteet

Liite A. Mitä on HPC?

High Performance Computing eli HPC-tietokonejärjestelmille on ominaista niiden nopea prosessointiteho, suorituskyykyiset sisäverkot ja suuri muistikapasiteetti, jotka mahdollistavat massiivisen määrän rinnakkaista käsittelyä. HPC-tietokonejärjestelmä on kokoelma tiiviisti toisiinsa liitettyjä tietokoneita, jotka toimivat rinnakkain yhtenä järjestelmänä suorittaakseen laajamittaisia laskentatehtäviä.

HPC-klusteri koostuu tyypillisesti monista yksittäisistä laskentasolmuista, joissa jokaisessa on yksi tai useampi prosessori, kiihdytin, muisti ja tallennustila. Solmut on yhdistetty toisiinsa huipputehokkaalla verkolla, jonka avulla ne voivat jakaa tietoja ja tehdä yhteistyötä tehtävissä. Lisäksi klusterissa on tyypillisesti erikoistuneita ohjelmistoja ja työkaluja resurssien hallintaan, kuten töiden aikatauluttamiseen, tietojen jakamiseen ja suorituskyyvyn seurantaan. Sovellusten nopeuttaminen onnistuu jakamalla tietoja ja jakamalla tehtäviä niin, että työ suoritetaan rinnakkain.

HPC-tietokonejärjestelmä voidaan toteuttaa esimerkiksi perinteisenä tietokoneklusterina tai pilvilaskentajärjestelmänä. Supertietokone on HPC-tietokonetyyppi, joka on äärimmilleen viety HPC-tietokonejärjestelmä ja joka tarjoaa ylivoimaisen laskentatehon ja -nopeuden jopa muihin HPC-järjestelmiin verrattuna.

Viime vuosina suurteholaskenta on kehittynyt simulointiin perustuvaan tieteelliseen tutkimukseen keskittyvästä välineestä kaksoisrooliin, jossa käytetään simulointia ja koneoppimista (Machine Learning, ML). Yksi keskeinen HPC:n ja ML:n kehitystä edistävä tekijä on grafiikkaprosessoritekniikan (GPU) kehitys. GPU:t ovat erikoistuneita tietokonesiruja, jotka on suunniteltu käsittelemään suuria tietomääriä rinnakkain, mikä tekee niistä ihanteellisia joihinkin HPC-ratkaisuihin, ja ne ovat tällä hetkellä ML/AI-laskentojen standardi.

Suursuorituskyykyinen tietojenkäsittely on tärkeää useista syistä:

- Nopeus ja tehokkuus: tutkijat ja insinöörit voivat ratkaista laajoja ongelmia, jotka eivät olisi mahdollisia tavanomaisilla laskentaresursseilla.
- Tieteellinen löytäminen: HPC on ratkaisevan tärkeää monilla tieteenaloilla, kuten ilmastomallinnuksessa, molekyyliidynamiikassa ja laskennallisessa nestedynamiikassa. Sen avulla tutkijat voivat simuloida monimutkaisia järjestelmiä ja prosesseja, mikä johtaa uusiin oivalluksiin ja löytöihin.
- Tuotteiden suunnittelu ja optimointi: HPC:tä käytetään laajalti esimerkiksi ilmailu- ja avaruusteollisuudessa, autoteollisuudessa ja energia-alalla tuotteiden, prosessien ja materiaalien suunnittelun simulointiin ja optimointiin, mikä parantaa niiden suorituskyykyä ja lyhentää kehitysaikaa.
- Data-analyysi: HPC on olennaisen tärkeää myös suurten tietokokonaisuuksien, kuten havaintotutkimusten, simulaatioiden tai kokeiden tuottamien tietokokonaisuuksien analysoinnissa. Sen avulla tutkijat voivat tunnistaa tiedoista kuvioita ja korrelaatioita, joita olisi vaikea havaita perinteisillä laskentaresursseilla.

- Terveydenhuolto: HPC:tä käytetään terveydenhuollossa yhä useammin uusien hoitojen ja terapioiden kehittämiseen, mukaan lukien yksilöllistetty lääketiede, lääkkeiden löytäminen ja molekyylihallinnus.

Suurteholaskenta toimii yhdistämällä useiden tietokoneiden laskentatehoa sellaisten laajamittaisten tehtävien suorittamiseksi, jotka olisivat mahdottomia toteuttaa yhdellä koneella:

1. Klusterikokoonpano: HPC-klusteri koostuu useista sadoista tai tuhansista tietokoneista eli solmuista, jotka on yhdistetty erityisen nopealla verkolla. Jokaisessa solmussa on yksi tai useampi prosessori, muisti ja tallennustila.
2. Tehtävien rinnakkaistaminen: Laskentatyö jaetaan pienempiin, itsenäisiin tehtäviin, joita voidaan suorittaa samanaikaisesti klusterin eri solmuissa. Tätä kutsutaan tehtävien rinnakkaistamiseksi.
3. Tietojen jakaminen: Laskennassa tarvittava data jaetaan solmujen kesken siten, että jokaisella solmulla on osa datasta työstettävänä.
4. Laskenta: Kukin solmu suorittaa oman osuutensa laskennasta rinnakkain, ja tulokset jaetaan ja lopulta yhdistetään, kunnes työ etenee loppuun.
5. Seuranta ja valvonta: Klusterissa on ohjelmistotyökaluja, jotka valvovat solmujen suorituskykyä ja ohjaavat tehtävien ja tietojen jakamista. Näin varmistetaan, että laskenta toimii tehokkaasti ja tuloksellisesti.
6. Tuotos: Lopullinen tuotos on klusterin kaikkien solmujen suorittaman laskennan tulos. Tulos tallennetaan yleensä suureen rinnakkaiseen tiedostojärjestelmään ja/tai renderöidään graafisesti kuviksi tai muiksi visuaalisiksi esityksiksi, mikä helpottaa löytämistä, ymmärtämistä ja viestintää.

Hyödyntämällä monien tietokoneiden kollektiivista tehoa HPC mahdollistaa laajamittaisten simulaatioiden, data-analyysien ja muiden laskentaintensiivisten tehtävien suorittamisen murto-osassa ajasta, joka kuluu yksittäisellä koneella.

Liite B. Käyttötapaukset

Käyttötapaus 1: HY (ML, monikerrosanalyysi)

Regarding the methods/computing resources, please find the brief description and our questions as below (they do look super computational cost, and that's exactly why we plan to apply for multiple grants to cover the potential cost):

We are preparing a computationally intensive research project involving multiple AI modeling phases and large-scale data processing.

Project Overview:

- Data Scale: 5 million individual records, each with approximately 8,000 features, spanning 30-50 years.
- **Computational Tasks:**
 1. CNN Training on 5 million images (each 100 × 30-50 pixels) using multiple convolutional kernels to predict all disease outcomes.
 2. Large Language Model (LLaMA 3.8B / TabLLM) fine-tuning for feature prediction across all disease outcomes.
 3. Super Learner Ensemble consisting of 10 models predicting all disease outcomes.
 4. Model Fusion Strategies (early, medium, late) combining insights from all three modeling streams.

We estimate needing six to eight GPUs simultaneously over several weeks, with approximately 2–3 months of cumulative GPU compute time, in addition to a CPU-intensive preprocessing phase without GPU. Could you please advise on the following?

1. Availability of GPU Resources

- What GPU models are currently available on CSC platforms (e.g., A100, V100)?
- Are there partitions or queues that support long-running jobs using 4–8 GPUs concurrently?

2. Usage Constraints & Allocation Procedures

- What are the typical maximum GPU hours or node allocations per user/project?
 - Are there any per-user or per-group time limits?
 - Is it possible to separate CPU time and GPU time between different project phases?

3. Supported AI Frameworks & Data Workflow

- Which AI/ML frameworks are supported (e.g., PyTorch, TensorFlow)?
- Are there recommended methods for efficient large-scale data ingestion and preprocessing on CSC systems?

4. Scheduling & Planning

- What are the estimated queue times for multi-GPU jobs?
- Do you have any recommendations or best practices to ensure efficient scheduling and utilization across the different phases?

Käyttötapaus 2: Tilastokeskus I (Tekoälyavusteinen ohjelmointi)

Koodiapuri

FIONA-etäkäyttöjärjestelmän sisällä toimiva LLM-pohjainen ohjelmoinnin tuki.

Tähän ajatukseen liittyvä selvitystyö on osana FIRE-hanketta.

Suurteholaskenta toisi tehokkuutta ja nopeutta

Käyttötapaus 3: Tilastokeskus II (Agenttiympäristö)

Agenttipohjainen analyysi

FIONA-etäkäyttöjärjestelmän sisällä toimiva työkalu, joka mahdollistaisi tutkijoille agenttien hyödyntämisen analyysien teossa.

Tietoturvasyistä agenttien hyödyntämiseen FIONAssa suhtaudutaan erittäin varovaisesti.

Käyttötapaus 4: Tilastokeskus III (ML-koulutus)

Koneoppimismallien kouluttaminen FIONA-etäkäyttöjärjestelmän sisällä
Kaivannee suurteholaskentaa.

Käyttötapaus 5: Aalto I (Etäpilvipalvelu)

API-pohjainen laskenta esimerkkinä verda.com jonka rautaa löytyy Pitäjänmäeltä

Käyttötapaus 6: Aalto II (Kubernetes HPC:ssä)

Kubernetes-klusterin laajennus esimerkiksi LUMI AI Factoryyn

Käyttötapaus 7: Aalto III (Paikallisten LLM:n käyttö)

Paikalliset kielimallit Ollama & Open WebUI Tämä on saatu toimimaan SAML-autentikoinnilla

Käyttötapaus 8: Aalto IV (Dedikoitu GPU)

Virtuaalikone, johon saa suoraan yhden tai useamman GPU:n kokonaan tutkijan käyttöön

Käyttötapaus 9: Findata I (Lähestymistapa laajan laskentakapasiteetin käyttöön)

Laajan kuva-aineiston analysointi tai genomitiedon, joihin ei meinaa nykyisen ympäristön paukut riittää. Tähän yhdistetään käyttäjien luvitus ja laskutus, onko akateeminen tutkimus vai kaupallinen, ja se kuinka anonyymisyys varmistetaan (mallit, tulokset etc). Rakennetaanko API laskentaan vai pitääkö datan uida ensin CSC:lle talteen. Kuinka järkevää on tehdä data siilo CSC:lle Lumen kylkeen?

Käyttötapaus 10: Findata II (Laskentakapasiteetin laajennus, kuva, pseudonymisaatio)

Findatan täytyy pseudottaa aineistot (ml isot kuva-aineistot) sekä pyörittää NER-mallia tekstiaineistojen käsittelyssä, tähän tarvittaisiin lisäpaukkuja

Käyttötapaus 11: Findata III (HPC toisilain alla)

Federoitu laskenta/oppiminen, kuinka mallit saadaan uimaan HPC laskentaan ja sieltä ulos toisilain kirjainta noudattaen, tuoko EHDS omia mausteitaan

Käyttötapaus 12: Varha I (Federoitu laskenta, histopatologia, OMOP)

AI-driven histopathology and OMOP-based modeling for metastatic risk in prostate cancer

Introduction

Prostate cancer is common in older men, with about 5,000 new cases diagnosed annually in Finland. Although overall prognosis is favorable, metastatic progression still causes roughly 1,000 deaths each year. Early identification of high-risk patients is critical, yet traditional risk factors have limited prognostic power. Emerging AI analyses of diagnostic biopsy slides show promise for predicting metastasis, but most studies remain small and lack long-term follow-up.

Rationale

Finland's integrated care enables comprehensive, longitudinal datasets linking diagnosis, treatment, and follow-up—ideal for developing and validating robust image-based riskmodels.

Materials and methods

To predict metastatic progression in prostate cancer patients, we will:

1. Develop deep learning models trained on digitized prostate biopsy images.
2. Determine metastasis onset dates from unstructured radiology reports stored in the OMOP CDM using validated LLM-based NLP methods.

Current status

1. The Ruusu vuori group has developed a robust research infrastructure for whole-slide image analysis.
2. The Ettala–Nieminen team is finalizing an LLM-based model to extract metastasis onset from unstructured radiology reports, with clinical validation underway.
3. Auria Clinical Informatics established the Varha OMOP CDM database in 2021 and continues to maintain and expand it.

The need

1. Scaling requires high-performance compute and storage (GPU-accelerated training for whole-slide images and LLM-based NLP).
2. Once the infrastructure is established with Varha data, we will expand to HUS and TAYS histopathology and OMOP datasets to build a privacy-preserving, federated multicenter image-analysis network.

Käyttötapaus 13: Varha II (Tekstiaineiston anonymisoinnin parantaminen suurten kielimallien avulla)

Tekstiaineiston anonymisoinnin parantaminen suurten kielimallien avulla, esim. maskaamalla potilaskertomustekstistä (NER, named entity recognition) tai sosiaalihuollon kertomuksista tutkimuskysymykseen liittymättömiä nimiä, henkiötunnuksia ja puhelinnumeroita. Tässä tarvitaan GPU-laskentaa avointen painojen kielimallien avulla (esim. gpt-oss:120b) ja aineistona hyvinvointialueen tietoaaltaan kertomustekstit (osajoukot ovat helposti miljoonia dokumentteja). Ajoon tarvitaan

tyypillisesti jokunen Nvidia H100 GPU-kortti ja menetelmästä riippuen, jonkin verran aikaa (päiviä tai viikkoja).

Käyttötapaus 14: Varha III (Raskaiden kielimallien käyttö kevyempien kielimallien hienosäädössä)

Oletetaan, että potilaskertomustekstistä pitää etsiä kategorisia muuttujia, kuten kirurgisten haattatapahtumien vakavuusluokka, potilaan tupakointistatus tai vain rakenteistaa vapaat tekstidokumentit siten, että kertomuksiin kirjatut tiedot saadaan omaan numeeriseen sarakkeeseensa. Tämän onnistuu jo nykyään raskaiden ja kyvykkäiden kielimallien avulla, mutta useiden tutkimus- ja kehityshankkeiden tarkoitus olisi päästä pidemmälle: tarvitaan kevyt ja tehokas täsmämalli, jota voidaan helposti ajaa osana potilaskertomusjärjestelmää esimerkiksi paikallisessa konesalissa. Tämä tuo "suvereniteettia", "tietoturvaa" ja "energiatehokkuutta" eli edustaa "kestävää kehitystä" ja tuottaa "eurooppalaisen ratkaisun", jos pelkkä arkijärjen mukainen fiksu toiminta ja IT-budjetin säästäminen ei kelpaisi tiedehankkeen rahoittajalle perusteluksi. Pienen mallin hienosäätöä varten potilaskertomustietoa kannattaa luokitella raskaiden ja kyvykkäiden mallien avulla. Näin pienempien kielimallien siirto-opettamiseksi saadaan erinomaisesti annotoitua dataa sekä kuhunkin luokitteluun liittyvät päättelyketjut (ns. ison kielimallin tuottama "thinking"-osio).