Aalto University
School of Science

Isaacus -hankkeen tietoallasratkaisujen arviointi tutkimuskäyttössä

Richard Darst & Mikko Hakala & Kimmo Kaski
Tietotekniikan laitos
Aalto -yliopisto

Selvityksen tavoite

Tässä selvityksessä arvioitiin Aalto-yliopiston toimesta Helsingin ja Uudenmaan sairaanhoitopiirin (HUS) ja Varsinais-Suomen sairaanhoitopiirin (VSSHP) tietoallas-ratkaisujen toteutuksia. Ensisijaisesti selvityksessä on arvioitu, kuinka hyvin ratkaisut toimivat tutkimuskäytön näkökulmasta ja mihin kokonaisuuksiin tulisi kiinnittää erityistä huomiota tämän käytön kehittämiseksi. Tarkastelun ulkopuolelle on sovitusti jätetty järjestelmien sisäinen datanhallinta, raportointi (business intelligence) ja datan soveltaminen organisaatioiden sisäiseen käyttöön.

Arviointi on toteutettu haastattelemalla maalis-huhtikuussa 2017 tahoja, jotka vastaavat näiden tietoallas-ratkaisujen toteuttamisesta. Selvitystä tehtäessä ratkaisuja oltiin edelleen kehittämässä ja erityisesti ratkaisujen integrointia eri Isaacus-partnereiden välillä ei oltu vielä aloitettu.

Organisaatiot

Sitra vastaa Isaacus projektin koordinoinnista ja rahoituksesta vuoden 2017 loppuun asti. HUS vastaa oman tietoaltaansa kehityksestä, jota koordinoi paikallinen asiantuntijaryhmä ja varsinaisesta tietoaltaan toteutuksesta vastaa Tieto Oy. VSSHP omaa sisäisen asiantuntijaryhmän, nimeltä kliininen tietopalvelu (KTP), joka antaa tukea data-analytiikalle ja vastaa paikallisen tietoaltaan suunnittelusta. VSSHP:n tietoaltaan implementaatiosta ja operoinnista vastaa Medbit Oy.

Lisäksi Sitra ja HUS & Tieto ovat informoineet, että Isaacus hankkeen osana Pohjois-Savon sairaanhoitopiiriin (PSSHP) Kuopio kehittää tietoallas-ratkaisua, jonka toteutuksesta – HUS:in tavoin – vastaa Tieto Oy. Tämän perusteella on odotettavissa, että PSSHP:n tietoallas on totetukseltaan identtinen HUS:in tietoaltaan kanssa, vaikka hallinnollisissa ratkaisuissa saattaa esiintyä pieniä eroavuuksia.

Nykytila

Selvityksen kirjoitushetkellä monet ratkaisut olivat vielä kehitysvaiheessa ja tutkijat eivät ole ehtineet testaamaan ratkaisuja tuotannossa. Selvitystä varten on arvioitu valmiilta osin jo pystytettyjä ratkaisuja täydennettynä sen hetkisillä suunnitelmilla. Teknisten ratkaisujen osalta arvioimme kokonaisuuden toimivan hyvin tutkimuksen tarpeisiin. Lisäksi organisaatioiden henkilöstön oma osaamistaso on erinomaista järjestelmien tukemiseen ja kehittämiseen.

VSSHP:lla on oma pieni yksikkö, joka aktiivisesti tarjoaa pieniä data-aineistoja paikalliselle tutkimukselle. Palvelu sisältää datan kuratoinnin ja analytiikkapalvelua. Lisäksi VSSHP:llä on alustava tuotantoympäristö, jossa voidaan tukea suurempia aineistoja ja raskaampaa analytiikkaa. Järjestelmä on kuitenkin oleellisesti prototyyppi ja ei välttämättä skaalaudu laajamittaiseen usean samanaikaisen tutkijan käyttöön.

HUS puolestaan pääosin odottaa oman lopullisen ratkaisunsa valmistumista. Tämä arvioidaan valmistuvan vuoden 2017 toisella puoliskolla. Tämän jälkeen HUS:n tarkoitus on evaluoida käytön perusteella, mihin suuntaan tutkimukselle suunnattua palvelukokonaisuutta kehitetään.

Tärkeimmät kehityskohteet

Kun tarkastellaan palveluita koko akateemisen tutkimusprosessin näkökulmasta, havaitsimme, että seuraavia osa-alueita ei ole määritelty riittävän tarkasti. Nämä kokonaisuudet ovat erittäin tärkeitä, jotta mahdollistetaan laaja-alainen tutkimuskäyttö.

1. **Löydettävyys.** Olemassa olevien aineistojen löydettävyys on hyvin oleellista määritettäessä tutkimuskysymystä. Mikäli data-aineistoa ei löydetä tai edes sen olemassaolosta ei ole tietoa, jää tutkimus suorittamatta. Tässä vaiheessa ei ole kuitenkaan olemassa katalogia tai muuta materiaalia, josta ilmenisi sekä yleisellä tasolla tietoa aineistoista, että myöhempää tutkimusta varten tarkempi aineisto kohtainen metadata. Hakupalvelu on suunnitteilla osana Isaacus hanketta, mutta on epäselvää toteutuuko tämä aikataulussa ja kuinka hyvin ratkaisu käytännössä soveltuu aineistojen löytämiseen. Metadatamalleista ja niiden integroinnista yli eri data lake –ratkaisujen ei ole myöskään tarkkaa tietoa tässä vaiheessa. Aineistojen esilletuomisen tärkeyttä ei tulisi aliarvioida, sillä tutkimuksen puolella on monia mahdollisia tapoja edetä. Mikäli Isaacus ja terveysdata haluaa tutkijoiden huomion, tulee tässä toimia aktiivisesti tutkimuksen suuntaan.
2. **Pääsyssä dataan** on myös mahdollisia pullonkauloja. Kaikissa tapauksissa tutkimusluvan myöntää erillinen hallinnollinen taho (eettinen toimikunta, THL, jne.). Mikäli lupia myönnetään vain näiden tahojen omien intressien priorisoinnin mukaisesti, rajoitetaan mahdollisia uusia tutkimusavauksia. Mikäli lupia myöntävät tahot keskittyisivät riskien minimointiin eivätkä suhtautuisi myönteisesti uusien mahdollisuuksien luomiseen, jäisi tutkimus tekemättä. Edelleen, mikäli syntyisi mielikuva, että lupien saanti olisi hankalaa, johtaisi tämä matalaan hakumäärään ja keskittäisi tutkimusta erityisesti yksiköiden sisälle.
3. **Rahoitusmalli.** Ne alueet, jotka käyttävät pilvipohjaista ratkaisua, voivat skaalata analytiikkaresursseja lähes rajatta. Avoimeksi jää, kuinka tämä kustannetaan. Tutkimuksessa raha on erittäin niukkaa ja tutkijat priorisoivat hankkeita, joissa voivat hyödyntää olemassa olevia resursseja. Mikäli käytön kulut kohdennettaisiin täysimääräisesti tutkimushankkeille, rajataan merkittävästi potentiaalista käyttäjäjoukkoa.
4. **Tekninen tuki**. Molemmissa tarkastelluissa tietoallasratkaisuissa ensisijainen ratkaisu intensiiviselle tutkimukselle ovat ns. *virtuaalikoneet* varustettuina analytiikkatyökaluilla. Useilla tutkijoilla on kuitenkin rajattu tekninen osaamistausta. Avoimeksi kysymykseksi jää, kuka tukee tutkijoita näiden työkalujen käytössä ja mahdollistaa sellaisen ympäristön, jota laaja tutkijakunta pystyy hyödyntämään.

VSSHP:llä tästä vastaa pieni paikallinen tukiorganisaatio, mutta tämän joukon ensisijaiseen käyttäjäkuntaan kuuluvat paikalliset lääkärit ja heidän analytiikkatarpeensa. Avoimeksi jää, kuinka laajaa ulkopuolista tutkijakuntaa tämä joukko pystyy palvelemaan. HUS:n osalta tämä kysymys on vielä kokonaan avoinna.

5. **Pitkän aikavälin kysymykset**, kuten integriteetti ja elinkaaren loppu. Koska näitä ratkaisuja ylläpidetään itse, tulisi data olla tulevaisuudessa siirrettävissä toisaalle tarpeen vaatiessa. Kuitenkin, pitkäaikaiseen rahoitusmalliin ja datan integriteettiin ei ole vielä olemassa vastausta.
6. Huomioitavaa on myös, että olemme evaluoimassa teknistä ratkaisua. Moniin vastaamattomiin hallinnollisiin kysymyksiin vastauksen antavat ylemmän tason päättäjät organisaatioissa, eivät nyt haastatellut kehittäjät tai tukihenkilöstö. Tällä hetkellä kokonaishanketta ohjaa Sitra, mutta mikä taho ottaa vastuun, kun Sitra katsoon tehtävänsä tulleen hoidettua?

Yleisiä havaintoja ja huomioonotettavaa

- Käyttökohteita on ajateltu olevan sekä pieni data että intensiivinen analytiikka myös isommalle aineistolle. Työkaluina voivat olla yksinkertaisessa tapauksessa Microsoft Excel ja intensiivisessä käytössä virtuaaliset koneet soveltuvilla analytiikkatyökaluilla.
- Molemmat evaluoidut tietoaltaat käyttävät Hadoop teknologiaa taustalla, joka soveltuu tällaiseen käyttöön hyvin. VSSHP (Medbit) toteuttaa ratkaisunsa siten, että data pysyy Suomessa. HUS (Tieto) puolestaan pohjaa ratkaisunsa Azuren julkiseen pilveen, jossa data sijaitsee pohjois-Euroopassa.
- Raskaaseen analytiikkaan VSSHP tarjoaa paikallisesti toteutettuja virtuaalikoneita. HUS puolestaan pilvipohjaista virtuaalikoneratkaisua. Pilvipohjainen ratkaisu skaalautuu helposti moniin eri käyttötapauksiin, mutta tässä tulee ottaa huomioon yllä mainitut tärkeimmät kehityskohteet.
- Kevyeen data-analyysiin (Excel) VSSHP tarjoaa etäkäyttötyöpöytiä, joissa dataa voidaan käsitellä VSSHP täysin hallinnoimassa ympäristössä. Tämä ratkaisu on tutkimuskäyttöön hyvin soveltuva ja joustava. Tällä hetkellä HUS:n suunnitelma on tarjota ratkaisua, jossa pieni data (pseudoanonymisoitu) siirtyisi tietoaltaasta tutkijan omalle työasemalle. Tämä on melko yllättävä, sillä tässä kohtaa HUS:ssa dataa ei pyritä hallinnollisesti pitämään vain suljetussa ympäristössä. Tämä edellyttänee jatkokeskustelua HUS:n kanssa.
- Data virtaa määritellysti yksiköiden omista lähteistä altaaseen, jossa se automaattisesti anonymisoidaan ja standardoidaan. Tutkimusluvan saatuaan tutkijoille tarjotaan pääsy anonymisoituun aineistoon joko rajapinnan kautta tai kopioimalla aineisto analytiikkapuolelle, jossa analyysi suoritetaan.
- Aineiston standardointi on oleellista, jotta varmistetaan datan tehokas hyödyntäminen ja linkitys toisiin aineistoihin. Evaluoidut organisaatiot ovat aktiivisia tämän kehittämisessä, mutta meillä ei ole tarkkaa tietoa nykytilasta. Lopputavoite on mahdollistaa datan linkittäminen yli eri tietoaltaiden.

- Dokumentaatio on kriittinen järjestelmien käytettävyyden osalta. Tämä on myös pidettävä ajan tasalla. VSSHP omaa sisäisen dokumentaatio, joka sisältää mm. käyttöoppaita. HUS on myös suunnitellut toteuttavansa dokumentaation. Käytännön toteutus on kuitenkin vielä avoinna.
- Keskitetty portaali tulee aikanaan hoitamaan lupahallinnan ja pääsyn avaamisen aineistoihin. Tämä ei kuitenkaan ole vielä valmiina ja on tässä vaiheessa epäselvää, kuinka automatisoitu ja nopea prosessi saadaan aikaiseksi. Alkuvaiheessa näyttää siltä, että aineistojen avaamisessa vaaditaan merkittävää manuaalista työtä paikallisilta tietoallasoperaattoreilta. Manuaalinen työ väistämättä lisää viivettä ja kustannuksia, joten pidemmällä tähtäimellä olisi mielestämme järkevää tehdä strateginen päätös, mikä on tavoiteltava automaation taso.
- Tällä hetkellä evaluoidut tietoaltaat eivät näe, että eri alueiden aineistoja linkitettäisiin toisiinsa nykyisissä järjestelmissä. Tällainen linkitys ja analyysi tultaisiin tekemään Tilastokeskuksen koordinoimassa ylemmän tason tietoallaympäristössä. Emme ole arvioineet näitä suunnitelmia, mutta toteutuessaan tämän järjestelmä olisi parhaimmillaan käyttöominaisuuksiltaan mahdollisimman vastaava nykyisiin ratkaisuihin verrattuna ja mahdollistaisi tehokkaan analytiikan sekä pienelle että suurelle aineistolle.
- Järjestelmien jatkokehitys vaatii palautetta todellisilta käyttäjiltä. Tässä on järkevää olla ennalta suunniteltu iteratiivinen prosessi, jossa palautetta kerätään systemaattisesti.
- Nykyisten organisaatioiden tukihenkilöiden asenne palveluun ja datan hyödyntämiseen on erinomainen. Kuitenkin henkilöstön on määrältään erittäin ohut ja ei mahdollista usean samanaikaisen hankkeen palvelua.
- Tietoturva- ja tietosuorajamallit ovat arvion mukaan hyvällä tasolla. *Aineistojen kohdalla on kuitenkin syytä huomioida, että mikäli aineisto siirtyy pois tietoaltaasta, se pysyy poissa.* Paras tapa vähentää tätä riskiä on tehdä tietoaltaasta mahdollisimman helppokäyttöinen analytiikan osalta, jolloin ei ole tarvetta käytännön syistä siirtää dataa ulkopuolelle. Tähän liittyen teimme kaksi yllättävää havaintoa: HUS:n aineistoja ei säilytetä Suomessa, vaan Microsoftin tarjoamassa pilvipalvelussa (EU:n sisällä). Lisäksi HUS:n suunnitelmissa pienimuotoinen anonymisoitu data voisi siirtyä tietoaltaasta suoraan tutkijoiden omille työlaitteille.
- Mitattavuus on laajempi kysymys, joka ei suoraan kosketa aktiivista tutkimusta. Kuitenkin rahoittajien puolelta lienee tärkeää mitata jollain tasolla käyttöä ja siitä saatavaa impaktia. Miten analytiikan onnistumista ja siihen käytettyjen investointien suuruutta mitataan ja millä kriteereillä järjestelmien tuleva rahoitus päätetään?

Yhteenveto

Olemme positiivisesti yllättyneitä nykyisestä kehityksestä ja suunnitelmista arvioitujen tietoaltaiden osana toimiviin analytiikkaympäristöihin. Kehityksen parissa toimivat henkilöt ovat erittäin motivoituneita ja osaavia alallaan. Kehitettävät alustat on suunniteltu huolellisesti ja ne vaikuttavat soveltuvan hyvin sekä pienempiin, että haastavampiin analyyseihin. Alustat pohjautuvat avoimiin ratkaisuihin, joka mahdollistaa pitkäaikaisen

tuen ja kehityksen. Esille nostamistamme tärkeistä kehityskohteista ollaan myös tietoisia kehittäjien puolella.

Kuitenkin, kuten monissa isoissa projekteissa, tässäkin on merkittäviä haasteita kommunikaatiossa ja johtamisessa. Erityisesti kansallisen tason suunnitelmat ja niitä edellyttävät eri osapuolten väliset integraatiot eivät ole vielä materialisoituneet. Sitra on tähän asti koordinoinut Isaacus-hanketta kokonaisuutena ja ohjannut kansallisen tason arkkitehtuurin edistymistä, mutta on siirtymässä syrjään hankkeesta vuoden lopussa. On epäselvää, mikä tahon ottaa vastuu koordinaatiosta tämän jälkeen ja varmistaa, että eri Isaacus-toimijat työskentelevät jatkossakin yhteisten päämäärien ja arvojen mukaisesti. *Tämä on niin keskeinen asiakokonaisuus hankkeen onnistumisen kannalta, että se tulee harkita ja toteuttaa erityisen huolella.*

Vaikuttaa selvältä, että eri toimijoiden sisäiset käyttäjät tulevat hyötymään rakennettavista tietoaltaista ja niihin liittyvistä analytiikkapalveluista. Ulkoiset (akateemiset, kaupalliset) käyttäjät hyötyisivät pääsystä aineistoihin, mutta tässä kohtaa kehitykseen on kiinnitettävä erityistä huomiota, jotta käytettävyys ja kustannusmalli pysyvät kilpailukykyisinä vaihtoehtoina.

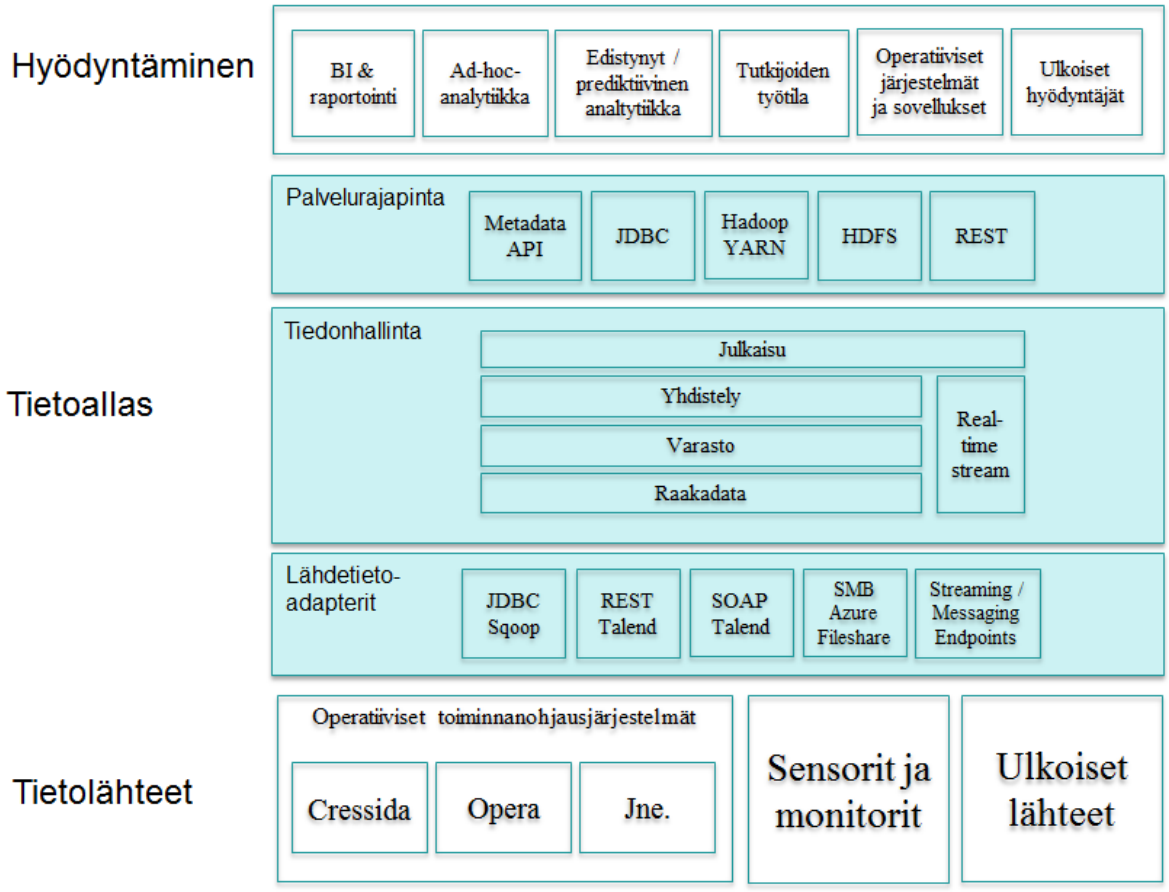
Lopuksi mainittakoon, että tällaisten laajojen järjestelmien – kuten tietoaltaat - evaluointi on haastavaa, kun lopullinen kokonaisuus ei ole vielä valmiina ja yksittäisen komponentin haasteet toiminnallisuudessa tai käytettävyydessä saattavat heijastua koko järjestelmään. *Tästä syystä tietoaltaiden todellinen evaluaatio tapahtuu vasta järjestelmää käyttävien tutkijoiden kokemusten ja palautteen kautta*. Kannustamme Sitraa keräämään aktiivisesti ja pitkäjänteisesti tutkijoiden kokemuksia tietoaltaasta ja huolehtimaan resursoinnein, että jatkokehityksen ohjauksessa tämä palaute toimisi ohjenuorana mahdollisimman käytettävän järjestelmän ja lopputuloksen toteuttamiseksi.

Liitteet

Arvioidut organisaatiot

HUS - Helsinki

HUS:llä on sisäinen tiimi, jotka vastaava tietoaltaan ja siihen liittyvän analytiikkaympäristön koordinoinnista. Ryhmä on kerätty juurikin tätä tarkoitusta varten ja varsinaisesta teknisestä toteutuksesta vastaa Tieto Oy. Tieto käyttää toteutuksessa avoimeen lähdekoodiin pohjautuvia ratkaisuja ja pyrkii sisäisesti tuotteistamaan tällaisen tietoallaskonseptin pystytyksen ja tukipalvelut. Tietoallaratkaisun ylläpidosta pystyy jatkossa vastaamaan myös muu, kuin ratkaisun toteuttanut taho. Tässä raportissa on haastateltu sekä HUS:n ja Tieto Oy:n toteutuksesta vastaavia henkilöitä.

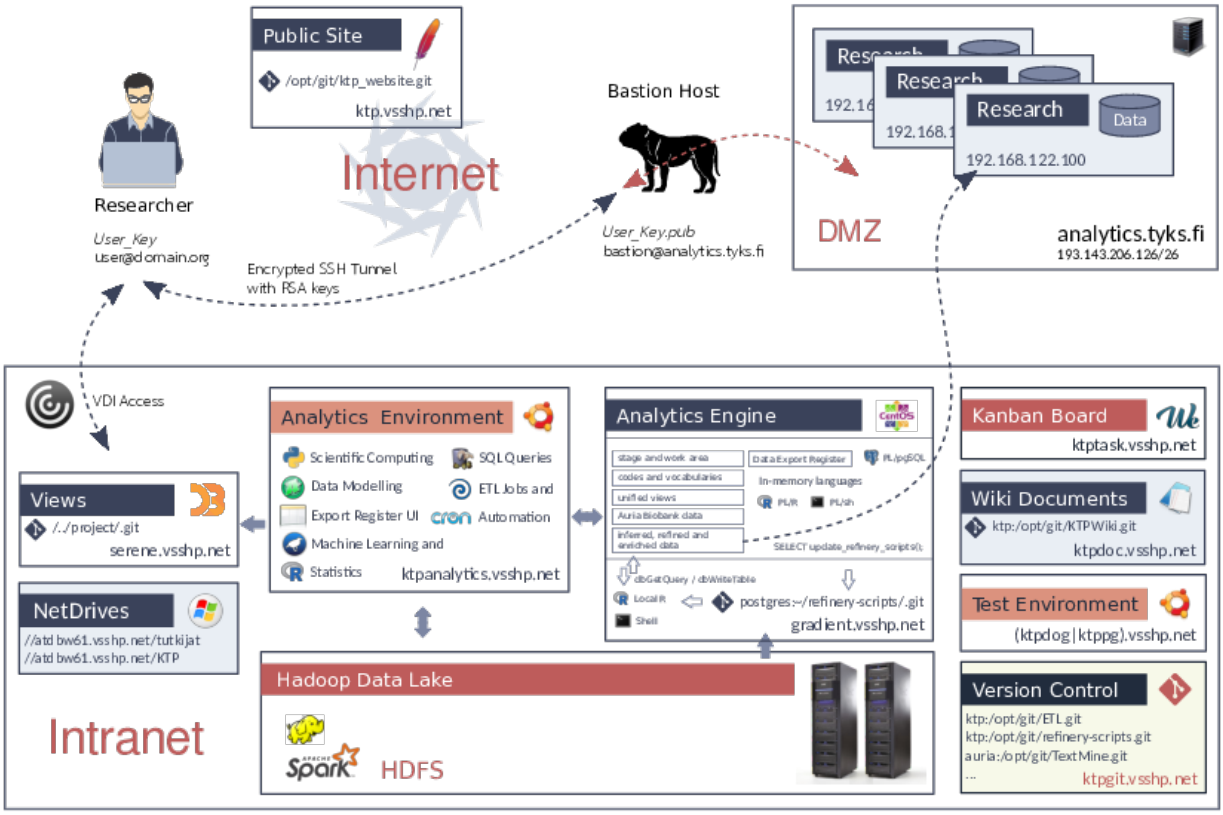


Kuva 1: *Skemaattinen kuva HUS:n tietoallasratkaisusta. Kuva keskittyy kuvaamaan datan virtausta tietolähteistä tietoaltaaseen ja edelleen loppukäyttäjille. Tutkijakunta on yksi osa loppukäyttäjistä.*

VSSHP - Turku

VSSHP:n sisällä toimii tekninen yksikkö nimeltä kliininen tietopalvelu (KTP), joka vastaa myös Isaacus hankkeesta ja sen koordinoinnista. Tämä yksikkö on kuuden hengen suuruinen ja toimii rajapintana datan ja kliinisen tutkimuksen välissä. Tärkeimmät palvelut ovat tuottaa analytiikkaa sairaalan sisäiseen käyttöön, datan koonti, yhdistäminen, anonymisointi ja pääsyn järjestäminen aineistoon.

VSSHP:lla ei ole KTP:n lisäksi muuta IT-henkilöstöä, vaan IT-palvelut tuottaa Medbit. Tämä on yhtiö, joka on alueen eri julkisten toimijoiden omistuksessa. Tietoaltaan osalta KTP on toteuttanut avoimeen lähdekoodiin pohjautuvan ensimmäisen generaation analytiikkaympäristön ja varsinaisen tietoaltaan tulee toteuttamaan Medbit.



Kuva 2: *Skemaattinen VSSHP:n analytiikkaratkaisusta. Kuva keskittyy kuvaamaan tutkijan pääsyä aineistoon (kuvaa eri asiaa kuin Kuva 1).*

PSSHP - Kuopio

Sitralta ja HUS & Tieto Oy:n tiedon perusteella PSSHP - Kuopio tekee yhteistyötä Tieto Oy:n kanssa rakentaakseen oman tietoallas ja analytiikka ratkaisunsa. Tämä perusteella on odotettavissa, että PSSHP – Kuopion ratkaisu tulee olemaan hyvin lähellä HUS:n ratkaisua ja kun

siitä osittain vastaavat samat henkilöt Tieto Oy:n puolella, jotka ovat rakentamassa HUS:n toteutusta. Koska PSSHP – Kuopion hanke on hyvin alkuvaiheessa ja vastaa HUS:n mallia, tässä selvityksessä on hyvin rajatusti aineistoa tähän hankkeeseen liittyen. Arviolta suurimmat erot näiden kahden mallin kohdalla ovat hallinnolliset ja käyttöpoliittiset erot.

Osa 1: Analytiikkaympäristöjen evaluointi

Analytiikkaympäristö tarjoaa paikan tutkimukselle hyödyntää tietoaltaiden data-aineistoja. Se yhdistää pääsynhallinnan, datan, laskentaresurssit ja tietoturvan. Tämä on erillinen ympäristö tietoaltaasta. Tutkimusnäkökulmasta analytiikkaympäristö on arvioinnin pääkohde.

Suunniteltu käyttäjäkunta

Tutkimuksessa tarpeet eroavat merkittävästi toisistaan. Tarpeet vaihtelevat pienen aineiston taulukkolaskennasta tutkimukseen, jossa kehitetään koneoppimista ”big data” aineistolle. Käyttäjien osaamisprofiilit voivat myös erota merkittävästi toisistaan.

HUS: Suunniteltu käyttäjäkunta sisältää talon sisäisen raportoinnin, lääkärit/lääketieteellisen puolen väitöskirjatutkijat ja laskennallisen tieteen. On mahdollista laajentaa käyttäjäkuntaa kaupallisiin toimijoihin. Tämän hetken rajaus tulee ennen kaikkea kyvykkyydestä tukea vain hyvin pientä käyttäjäkuntaa.

VSSHP: Suurin käyttäjäkunta on talon sisäiset lääkärit / tutkijat, erityisesti lääkärit jotka saavat taulukkolaskenta –tyyppistä raportointia KTP:n kautta. Analytiikkatarve on toistaiseksi varsin kevyttä. Mahdollisuus raskaampaan tietojenkäsittelyyn on erillisen ympäristön kautta, joskin tätä on toistaiseksi markkinoitu vain sisäisesti (ktp.vsshp.fi).

Yhteenveto: Molemmilla organisaatiolla on melko samanlaiset tavoitteet ja ratkaisuja, joita voidaan hyödyntää sekä pieneen, että raskaampaan analytiikkaan. Molemmissa tapauksissa rajoittava tekijä laajempaan käyttöön on tukihenkilöstön suuruus.

Ohjelmisto

Kaikki analytiikka tehdään käyttäen ohjelmistoja ja näiden saatavuus määrittelee, mitä voidaan analysoida. Ohjelmistojen osalta arvioimme kahta eri käyttötapausta 1) ”pienen data” tarpeet, eli tutkijat, joille riittää taulukkolaskenta ja valmiit yksittäiset ohjelmistopaketit ja 2) tieteellinen laskenta, eli tutkimus, jossa todennäköisesti toteutetaan analytiikka omalla ohjelmistolla ja ohessa kehitetään näitä menetelmiä.

HUS: Ohjelmistoympäristö on Microsoftin Azure pilvialusta. 1) Tarjolla on erillinen tietokantarajapinta (ODBC), jolla data voidaan tuoda ulos tietoaltaasta ja analytiikka voidaan täten suorittaa tutkijoiden omilla työasemilla. 2) Tutkijat saavat käyttöönsä Linux pohjaisen virtuaalikoneympäristön, jossa useimmin käytetyt työkalut ovat valmiiksi asennettuina. Lisäksi

tutkijat voivat itse edelleen muokata ympäristöä haluamakseen. Koska molemmat data ja laskenta tehdään pilvialustalla, on tässä enemmän mahdollisuuksia avata pääsy suoraan aineistoon ilman datan kopiointia.

VSSHP: 1) Windows etäkäyttötyöpöytä oleellisesti mahdollistaa kliinisen tutkimuksen käyttää Microsoft Excel taulukkolaskentaa tutkimaan pientä aineistoa. Tämä on yksinkertainen ja tietoturvallinen tapa, joka vastaa paikallisia tarpeita. Myös muuta standardia ohjelmistoa voidaan mahdollisesti asentaa etäkäyttötyöpöydille pyydettäessä. 2) Kuten HUS:n tapauksessa, erillinen virtuaalikonepohjainen Linux ympäristö voidaan pystyttää, jossa tutkijat voivat vapaasti muokata ympäristöä ja tehdä analytiikkaa. Tarvittava anonymisoitu aineisto kopioidaan tähän ympäristöön. Nykyisessä toteutuksessa on hyödynnetty tietokantapohjaista ratkaisua datojen tuomisessa ja tulevaisuudessa integraatio varsinaiseen tietoaltaaseen voidaan toteuttaa, kun lopullinen tietoallas on tuotannossa.

Yhteenveto: Sekä VSSHP:n Windows etätyöpöydät, että HUS:n pilvialustalla toimivat virtuaalikoneet ovat ideaaleja omaan käyttötarkoitukseensa. 1) VSSHP:n etätyöpöytä mahdollistaa analytiikan suoraan järjestelmässä, joka on tietosuojan kannalta oleellista. Tätä on hyvä verrata HUS:n ratkaisuun, jossa data kopioidaan ulos organisaatiosta. 2) Kehittyneempään käyttöön molemmat HUS ja VSSHP tarjoavat virtuaalikoneita, jotka mahdollistavat lähtökohtaisesti joustavuuden kaikkeen tutkimuskäyttöön. Periaatteessa HUS:n virtuaalikoneratkaisu on tehokkaampi ratkaisu, sillä se mahdollistaa monissa tilanteissa turvallisen pääsyn dataan ilman järjestelmän sisäistä kopiointia. Virtuaalikoneiden kohdalla on syytä muistaa, että nämä vaativat jonkin verran teknistä osaamista tutkimuksen puolelta.

Laitteisto

Laitteiston merkitys kevyessä analyysissä on pieni, mutta raskaampi analytiikka vaatii merkittävämpää suorituskykyä laitteistolta (tallennus, muisti, CPU, spesialisoitu laskenta GPU).

HUS: Analytiikkaympäristö sijaitsee Azuressa, jolloin resurssien saatavuus on lähtökohtaisesti helppoa. Konseptin mukaan maksetaan tarpeesta ja resurssit ovat välittömästi käytettävissä. Jotkin spesialisoidut ratkaisut (GPU) ovat saatavissa vain tietyissä maantieteellisissä alueissa. Tässä palvelumallisissa kustannusten kohdennus nousee kuitenkin esiin. Kuka maksaa resurssien käytöstä? Itse HUS tietoallaskonsepti on suunniteltu joustavaksi ja vastaava ympäristö voidaan asentaa myös muualle, kuin pilvialustalle.

VSSHP: Nykyinen ratkaisu käyttää omaa fyysistä laitteistoa, joka sijaitsee Medbitin konesalissa. Tämä mahdollistaa datan säilymisen Suomessa ja toisaalta laitteiston joustavan muokkauksen. Toisaalta, tämä tarkoittaa, että laitteiston ylläpitoon joudutaan investoimaan ylläpitoresursseja ja skaalautuvuudessa on useampia rajoituksia (konesalikapasiteetti, kilpailutukset). Tässä on syytä huomioida, että nykyinen ratkaisu on prototyyppi ja laitteistoalusta voidaan vaihtaa tulevaisuudessa.

Yhteenveto: Lyhyesti, HUS ja VSSHP ratkaisut eroavat erityisesti pitkäaikaisen suunnittelun ja nopean hyödynnettävyyden osalta. Mikäli pilvipohjainen ratkaisu on kustannustehokas ja vastaa organisaation tietosuojapolitiikkaa, on se todennäköisesti järkevin ratkaisu. Kuitenkin sen pystyttämiseen mene merkittävästi enemmän aikaa. Mikäli lopullinen käyttö jäisi vaatimattomaksi, voi yksinkertaisempi ratkaisua tuoda enemmän etua joustavuutensa kautta.

Datan tuonti analytiikkaan

Data tuodaan tietoaltaaseen useasta eri lähteestä, jonka jälkeen tämä virtaa tietoaltaassa muutamien prosessien lävitse ja lopulta aineisto on saatavilla analytiikkaympäristössä. Näiden prosessien nopeus ja helppokäyttöisyys vaikuttavat hyödynnettävyyteen.

HUS: Tällä hetkellä datan tuottavat tahot hoitavat integroinnin tietoaltaaseen. Tämä tuo raakadatan joka tämän jälkeen automaattisesti standardoidaan ja anonymisoidaan. Tämän jälkeen aineisto on valmis välitettäväksi eteenpäin eri käyttäjille. HUS:n ratkaisu sisältää myös soveltuvin osin aineiston metadatan keräämisen joka auttaa aineiston jatkokäsittelyssä.

VSSHP: Itse tietoallas on vielä kehitysvaiheessa, joten datan tuonnista ei ole vielä paljoa tiedossa. Tällä hetkellä KTP hoitaa aineiston siirron suoraan lähtöpaikasta ja edelleen prosessoinnin manuaalisesti loppukäyttäjälle.

Yhteenveto: Datan tuonti tapahtuu pääosin taustalla. HUS:n strategia on erittäin toimiva ja VSSHP:lla on käytännön kokemusta, miten tämä voidaan toteuttaa, kun varsinainen tietoallas saadaan tuotantoon.

Aineistojen linkittäminen

Data tuotetaan eri lähteissä ja tulee tietoaltaaseen useassa eri muodossa. Jotta dataa voidaan maksimaalisesti hyödyntää, tulee nämä aineistot linkittää toisiinsa.

HUS: Standardointiprosessi sisältää uniikin määrittelyn ja määritellyn salauksen. Tämä mahdollistaa, että kaikki aineisto on linkitettävissä toisiinsa. Eri sairaanhoitopiirit koordinoivat, että tämä uniikki määrittely olisi standardoitu, jotta pitkällä aikavälillä aineistot yli sairaanhoitopiirien voitaisiin linkittää toisiinsa.

VSSHP: KTP pystyy toteuttamaan datan linkittämisen eri lähteiden välillä, mutta tällä hetkellä tämä on manuaalinen prosessi. Tulevaisuudessa tämä on tarkoitus automatisoida ja tässä VSSHP ja HUS tekevät yhteistyötä.

Yhteenveto: Molemmissa organisaatioissa datan linkittäminen on huomioitu ja tässä pyritään standardoimaan kansallisen tason ratkaisua. Tärkeänä kysymyksenä tämän selvityksen ulkopuolelle jää, miten mahdollistetaan tutkimus linkitetyllä aineistolla. Nykyinen suunnitelma on toteuttaa tämän käyttäen Tilastokeskuksen toteuttamaan keskitettyä alustaa, mutta tämä ei ole vielä tuotannossa, eikä sitä ole arvioitu osana tätä selvitystä.

Lupahallinta

Jotta teknistä ratkaisua voidaan hyödyntää, tulee tutkijoiden saada pääsyhallinkautta käyttöoikeus. Tämä on ennen kaikkea hallinnollinen kysymys, mutta arvioimme tätä ja sen teknistä toteutusta.

HUS: Pääsy avataan siinä yhteydessä, kun tutkija saa eettisen komitean kautta luvan käyttää aineistoa. Pidemmällä aikavälillä tämä voi vaatia tarkastelua THL:n puolella, jotta voitaisiin yhdistää tutkimus ja aineistot automaattisemmin. Teknisesti tässä käytetään sairaalan omaa lupahallintaa tai Suomi.fi –palvelua, jota kautta saadaan vahva tunnistautuminen ulkopuolisille käyttäjille. Myös muita autentikointilähteitä olisi teknisesti mahdollista käyttää. Tällä hetkellä suunnitelmissa on autentikoida vain suomalaisessa järjestelmässä olevia henkilöitä.

VSSHP: Kuten HUS:n tapauksessa pääsy on sidottu sairaalan myöntämään tutkimuslupaan. Tällä hetkellä tunnistautuminen etätyöpöytiin käyttää sairaalan sisäistä järjestelmää ja virtuaaliset työasemat käyttävät avainpohjaista tunnistautumista, joka vaatii manuaalista työtä. Tämä tosin mahdollistaa pääsyn teknisesti kaikille ulkopuolisille käyttäjille.

Yhteenveto: Pääsynhallinta on ennen kaikkea hallinnollinen kysymys, mutta teknisesti tulisi olla sekä helppo tapa sisäisille käyttäjille, että joustava tapa ulkopuolisille käyttäjille päästä käsiksi aineistoihin. Ensin tulisi päättää ketkä ovat haluttu käyttäjäkunta (kansallinen vain kansainvälinen käyttö) ja suunnitella ratkaisu tätä vasten.

Pääsynhallinta

Kun tutkija on saanut luvan käyttää aineistoa, tulee aineisto kohdentaa käyttäjälle järjestelmässä. Ideaalissa tapauksessa tämä toimisi automaattisesti, mutta tällä hetkellä tämä vaihe sisältää vielä hieman manuaalista työtä ja toimii myös toisena tarkistuspisteenä luvan kohdalla.

HUS: Erillinen lupahallintapalvelu, johon myös tutkimusluvat kirjataan, pitää kirjaa käyttäjistä oikeuksista. Tämä sisältää tiedon mihin dataan käyttäjällä on oikeus ja järjestelmä pystyy avaamaan pääsyn tietoaltaassa tähän aineistoon. Tällä hetkellä prosessi on vielä manuaalinen ja automaatio riippuu siitä, kuinka strukturoiduksi koko prosessi pystytään kehittämään. Hallinta onnistuu sekä yksilö- että ryhmätasoilla.

VSSHP: Nyky-ympäristössä tämä on manuaalinen prosessi. Lupia voidaan myöntää joustavasti käyttäjän, ryhmän ym. perusteella.

Yhteenveto: Täysin automatisoitu prosessi olisi ideaalisin, sillä tämä toimisi nopeasti, tehokkaasti ja vähentäisi mahdollisia virheitä. Tämä vaatii kuitenkin taustalle hyvin strukturoidun prosessin. Kansallisesta lupaportaalista ei ole tässä vaiheessa tarkempaa tietoa, joten integraatiota tähän ei ole evaluoitu.

Dokumentaatio

Jotta järjestelmiä voidaan käyttää tehokkaasti pitkällä aikavälillä, tulee se ja sen käyttö olla dokumentoitu.

HUS: Tutkijoille ollaan kirjoittamassa käyttöopasta, mutta tämä ei ole vielä valmiina eikä avoimena.

VSSHP: Sisäinen wiki-alue toimii hallinnollisena dokumentaationa. Analytiikan käyttöön tämä voidaan toimittaa tutkijoille pyydettäessä. Dokumentaatio on myös luotu muotoon, jossa se voidaan helposti tulostaa kirjamuotoon.

Yhteenveto: Dokumentaatio on kriittinen tekijä minkä tahansa järjestelmän kohdalla. Se pitää myös pitää päivitettynä, jotta sen hyöty säilyy. Parhaimmillaan tässä voidaan tehdä standardisoitu eri tietoaltaiden ylitse ja mahdollisesti pitää aineisto täysin avoimena. Tällöin se tukee maksimaalisesti järjestelmien käyttöä.

Automaatio

Pitkän aikavälin resursoinnin ja ylläpidettävyyden, sekä tehokkaan käytön näkökulmasta kaikki mahdolliset välivaiheet aina datan tuonnista pääsynhallintaan olisi järkevä automatisoida.

HUS: Lopullinen järjestelmä parhaimmillaan mahdollistaa automaation datan tuonnin, muokkauksen ja anonymisoinnin osalta.

VSSHP: Tällä hetkellä KTP:n päätavoite on strukturoida olemassa oleva aineisto. He ovat mukana eri standardointia koskevissa hankkeissa. Kun lopullinen järjestelmä on olemassa, heillä on hyvät valmiudet laajaan automatisointiin.

Yhteenveto: Molemmissa organisaatiossa on tahtotilana automatisoida datan prosessointi ja he osallistuva aktiivisesti tähän kehitykseen. Kuitenkin laajempi automaatio kansallisella tasolla on suurempi tavoite, joka vaatii paljon työtä ja koordinaatiota. Nähtäväksi jää, pysyykö yhteinen visio ja saadaanko tätä kehitystä vietyä eteenpäin tulevaisuudessa.

Tietoturva

Järjestelmissä, joissa säilötään lääketieteellistä ja erityisesti henkilötietosuojan alla olevaa aineistoa, tulee noudattaa asianmukaisia tietoturvakäytäntöjä.

HUS: Tällä hetkellä suunnitelmissa on, että pienimuotoinen (anonymisoitu) aineisto lähtökohtaisesti poistuisi tietoaltaasta. Tutkijat yhdistäisivät ohjelmistonsa tietoaltaaseen ja hakisivat aineiston omalla työkoneelleen analyysia varten. Suurempi aineisto evaluoitaisiin virtuaalikoneilla ja aineisto pysyisi ympäristön sisällä.

VSSHP: Lähtökohta on, että data ei lainkaan poistu tietoaltaasta. Pieni aineisto evaluoidaan etätyöpöytiä hyödyntäen ja muu aineisto ympäristössä olevilla virtuaalikoneilla.

Yhteenveto: Molemmat organisaatiot noudattavat hyviä tietoturvakäytäntöjä. Käyttäjät pyritään tunnistamaan vahvan tunnistautumisen kautta ja kullekin annetaan rajattu pääsy vain luvanmukaiseen aineistoon. Ulkoiselle tutkimukselle annetaan pääsy vain anonymisoituun aineistoon. Riskien näkökulmasta nykyiset ratkaisut eivät pysty estämään aineiston viemistä pois järjestelmistä, joskin tämä joissain tapauksissa sopimuksen vastaista. Tästä huolimatta on yllättävää, että HUS:n kohdalla pieni aineisto on mahdollista suunnitellusti vielä tietoaltaan ulkopuolelle. Huomioitavaa on, että tässä raportissa arvioitiin lähinnä tietoturvaan liittyviä teknisiä toteutuksia. Esimerkiksi tietoturvaan liittyvät prosesseja ja niiden auditoinnit ovat tämän tarkastelun ulkopuolella.

Standardointi

Prosessit ja dataformaatit tulisi standardoida ja yhtenäistää mahdollisimman hyvin yli eri hoitopiirien (ja jopa laajemmin globaalisti, sikäli kun tämä on järkevää). Tämä helpottaa aineistojen käyttöä ja erityisesti mahdollistaa tulevaisuudessa aineistojen yhdistämisen.

HUS: Tieto Oy pyrkii kehittämään HUS ratkaisusta sellaista, jota voidaan käyttää laajemmin Suomessa. PSSHP - Kuopio on lähtenyt kehittämään omaa järjestelmäänsä tältä pohjalta.

VSSHP: KTP on mukana eri suomaisissa ja pohjoismaisissa aineiston standardointiin liittyvissä projekteissa. Heidän tekninen ratkaisunsa on erilainen HUS:in verrattuna, mutta ennen kaikkea siksi, että Medbit ei ole vielä toteuttanut varsinaista tietoallasta.

Yhteenveto: Molemmat organisaatiot näkevät standardoinnin tärkeäksi ja ovat aktiivisesti yhteydessä keskenään. Tässä vaiheessa monet asiat ovat kuitenkin vasta suunnitteilla ja varsinainen toteutus puuttuu.

Osa 2: Tietoallasratkaisujen evaluointi

Tietoallas on paikka, joka pitää varsinaisen raakadatan ja tarjoaa aineistoa edelleen analyysiympäristöön. Tutkijoilla ei ole tähän pääsyä, kun tutkimus tarvitsee dataa, se siirretään tai avataan erillinen rajapinta, joka mahdollistaa pääsyn kyseiseen aineistoon. Virtuaalikonepohjainen analytiikka saattaa sijaita samassa ympäristössä, mutta nämä ovat loogisesti erillään.

Tekniset yksityiskohdat

HUS: Tietoalla on toteutettu HDInsight ratkaisulla Microsoft Azure pilvialustalla. Ratkaisu ja aineistot sijaitsevat EU:ssa, mutta ei Suomen rajojen sisäpuolella. Azure mahdollistaa suoraan

analytiikkaan liittyviä työkaluja, kuten Spark -ohjelmisto. Tieto Oy kehittää konseptia siten, että se on monistettavissa myös muihin organisaatioihin.

VSSHP: Tietoalla pohjaa myös Hadoop teknologiaan, mutta käytettävä ratkaisu on Cloudera. Ratkaisu tulee sijaitsemaan Suomessa olevissa konesaleissa ja kokonaisuudesta vastaa Medbit.

Yhteenveto: Molemmat organisaatiot käyttävät samaa ratkaisua. Eri version käyttö ei aiheuttane haasteita integraatiolle tai yhteensopivuudelle.

Data lähteet

Jotta alusta olisi käytettävä, tulee eri lähteistä peräisin oleva integroida tähän. Tämä on alussa melko työläs prosessi, joka tulisi lopuksi toimia automatisoidusti.

HUS: Integrointi tapahtuu alkuperäisen datan tuottavien tahojen toimesta. Tavoite on integroida mahdollisimman monta datalähdettä osaksi tietoallasta.

VSSHP: KTP:llä on osaamista eri datalähteiden prosessoinnista, mutta Medbit on taho, joka toteuttaa lopullisen integraation. Medbit informoi, että jokainen datalähde tulee tarkistaa ja arvottaa ennen integrointia. Oma ymmärryksemme on, että asiakkaan (VSSHP/KTP) tulee pyytää näitä datalähteitä integroitavaksi. KTP:lla on alustava lista ensimmäisen vaiheen integraatioista.

Yhteenveto: Molemmilla organisaatioilla on suunnitelma datan integroinnista. HUS on optimistisempi erilaisten datalähteden osalta, kun Medbit puolestaan tekee integroinnin asiakkaiden pyynnöstä. Emme tiedä tarkemmin kuka päättää kuinka monta tai mitkä integraatiot tullaan toteuttamaan. Tästä huolimatta integraatio on jatkuva prosessi, jota tulee päivittää ja kehittää.

Dokumentaatio

Dokumentaatio on oleellinen sekä itse järjestelmästä, että sen sisältämän datan osalta. Dataan liittyvää metadataa voidaan edelleen hyödyntää datan automaattisessa prosessoinnissa.

HUS: Datan keräämisen ohella tullaan keräämään myös määriteltyä metadataa, jota hyödynnetään tietoaltaassa datan prosessoinnissa. Poislukien tutkimukselle ja käyttäjille suunnattu dokumentaatio, varsinaisesta tietoaltaan dokumentaatiosta ei keskusteltu.

VSSHP: Raportin kirjoitusvaiheessa tämä oli vielä avoin kysymys.

Yhteenveto: Dokumentaatio on usein viimeisin asia, mitä hankkeissa pohditaan. Tässä vaiheessa ei ole saatavilla aineistoa, jotta tätä voitaisiin arvioida.

Hyödynnettävyys

Yhden sairaanhoitopiirin toimiva konsepti on ideaalissa tapauksessa muokattavissa helposti toimivaksi toisaalle. Tällöin saadaan suoria kustannussäästöjä ja dataa ja osaamista voidaan vaihtaa alueiden välillä.

HUS: Ratkaisu pohjaa Tieto Oy:n konseptiin ja tästä pyritään tekemään helposti monistettavissa oleva ratkaisu eri alueiden käyttöön. Ratkaisu pohjaa avoimen lähdekoodin ohjelmistoihin ja pystytettävissä erilaisiin alustoihin, ei pelkästään Azuren päälle, joka on ratkaisu HUS:n kohdalla.

VSSHP: Medbit toteuttaa tietoallaratkaisun. Hyödynnettävyys muualla on vielä avoinna.

Yhteenveto: Molemmilla organisaatioilla on tavoitetilana tehdä laajemmin hyödynnettävä ratkaisu.

Laskenta

Laskenta lähtökohtaisesti tehdään erillisessä analytiikkaympäristössä. Mutta suurten aineistojen kohdalla voi olla perusteltua siirtää laskenta tehtäväksi suoraan tietoallasympäristössä. Tämä vaatisi tietoturvasyistä erityishuomiota ajettavalta koodilta ja olisi luontevaa evaluoida tapauskohtaisesti, mikäli tällaisia tarpeita ilmenee tulevaisuudessa.

HUS: Analytiikka tehdään lähtökohtaisesti tietoaltaan ulkopuolla, mutta teknisesti laskenta suoraan tietoaltaan puolella olisi mahdollista.

VSSHP: Kuten HUS.

Yhteenveto: Vastaus on alkuperäisin oletuksen mukainen. Mikäli olisi hyvin perusteltu tarve ja resurssit, tämä olisi teknisesti mahdollista. Tällainen tarve vaatisi aina erityisen käsittelyn.

Pitkän aikavälin suunnitelma

Nykyinen järjestelmä on vasta rakenteilla, mutta mitä tapahtuu tulevaisuudessa, kun järjestelmä pitää vaihtaa? Kuinka data on ajateltu siirtää toiseen järjestelmään ja pysyykö data muuttumattomana siihen asti?

HUS: Tämä on ajateltu vastuuttaa Tieto Oy:lle. Tieto on suunnitellut tekevänsä konseptista pitkäikäisen ja tarjoavansa sille tukea. Palvelun tarjoajana Tieto Oy olisi vastuussa siirtymisestä toiseen järjestelmään, kun se olisi ajankohtaista.

VSSHP: Medbit toteuttaa asiakkaan pyynnöt.

Yhteenveto: Järjestelmiä ylläpidetään läheisessä yhteistyössä asiakkaan kanssa. Täten toimittajan tulisi pystyä tekemään ratkaisuja tarpeiden pohjalta. Pitkän aikavälin ratkaisuista voisi

olla järkevää nostattaa hallinnollista keskustelua, sillä tämä ei ole oleellisesti tekninen kysymys. Emme huomanneet, että kummallakaan organisaatiolla olisi ollut suunnitelmia aineiston pitkäaikaisen integriteetin osalta.