HUS TIETOALLAS SUUNNITTELU

Versio 1.1, 8.12.2017

Johdanto

Tässä dokumentissa kuvattu suunnitelma koskee tietoallasta, jolla voidaan

- toteuttaa erilaisia lääketieteen tutkimushankkeita tarjoamalla tiedon tallennukseen, hallintaan ja analysointiin sopivia työkaluja, sekä
- tarjota tehokas alusta raportointikehitykeen, ja
- tukea tulevaisuudessa operatiivisen tuotannon tarpeita tarjoamalla taustajärjestelmiä älypalveluille.

Tietoallas on nk. Big Data – järjestelmä, joka kykenee käsittelemään suuria tietomääriä ja mahdollistaa suurten tietomäärien analytiikan.

Arkkitehtuurin tavoite on kehittää tietoallas, jossa seuraavat ominaisuudet on otettu tarkasti huomioon:

- **Joustavuus**. Kaikkia käyttötapauksia ei tuotannon suunnittelun yhteydessä tiedetä, ja alustan suunnittelun tärkein päämäärä on joustavuus. Joustavuudella taataan tulevien käyttötapausten suunnittelun ja käyttöönoton helppous.
- **Tietoturva**. Potilastietojen käsittely on lailla säädeltyä ja se asettaa vaatimuksia siihen kuinka sen käyttöä tulee rajata ja seurata.
- **Yhdistettävyys**. Palveluun voidaan helposti yhdistää uusia tietolähteitä sekä liittää analytiikkaan liittyviä työkaluja.
- **Saavutettavuus**. Palvelu tulee olla saavutettavissa ja käytettävissä aina. Palveluun tallennetut tiedot tulee olla suojattu katoamista vastaan.
- **Ylläpidettävyys**. Big data – järjestelmät kehittyvät nopeasti. Järjestelmä tulee olla ylläpidettävä ja se tulee olla mahdollista päivittää uudempaan helposti jotta uudet tiedonkäsittelymenetelmät voidaan tuoda tuotantoon helposti ja nopeasti.

Tausta ja tavoitteet

HUS Tietoaltaasta perustuu 2015-2016 toteutetulle hankkeelle HUS Big Data Platform, jossa suunniteltiin ja pilotoitiin HUS:n tulevaisuuden tiedonhallinnan kokonaisarkkitehtuuria. Hankkeen keskeisenä tuloksena syntyi suunnitelma ja pilottitoteutus uudentyyppisestä big data ja pilviteknologioita hyödyntävästä, keskitetystä tiedonhallinta-alustasta, jota alettiin kutsua HUS Tietoaltaaksi.

HUS Tietoallas on valittu myös yhdeksi Sitran käynnistämän Isaacus-hankekokonaisuuden yhteistyöhankkeeksi. Isaacus-hankekokonaisuuden tavoitteena on luoda kansallinen toimija kokoamaan ja koordinoimaan hyvinvointidataa kansallisella tasolla. Muodostettavaa kansallista toimijaa kutsutaan hyvinvoinnin palveluoperaattoriksi.

HUS Big Data Platform -hankkeessa asetetut tavoitteet uudelle arkkitehtuurille olivat:

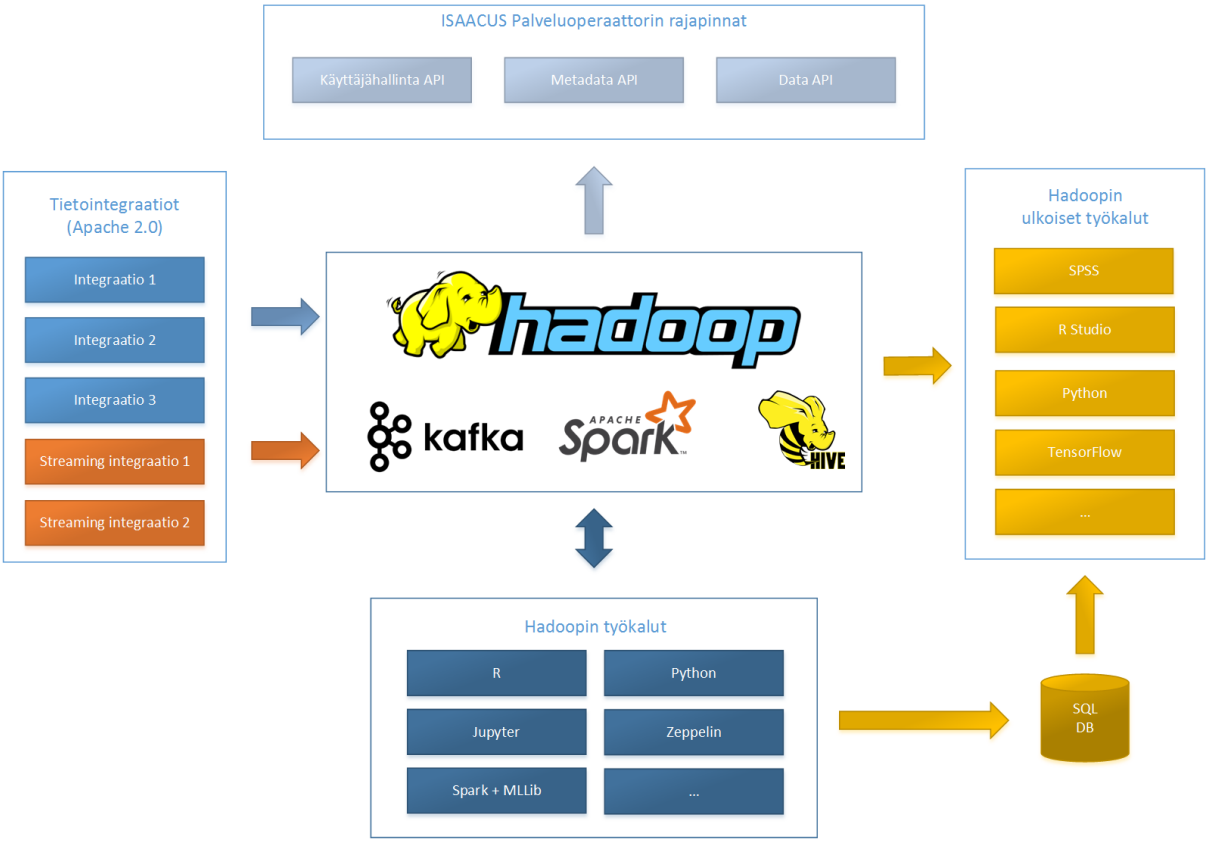
1) nopeuttaa raportointikehitystä
2) mahdollistaa tehokkaan analytiikan
3) helpottaa kliinistä raportointia ja mittareiden yhdenmukaisuutta
4) mahdollistaa yhdet luvut
5) antaa valmiudet Big Data -menetelmiin
6) tukee Apotin käyttöönottoa
7) helpottaa raportointityökalujen vaihdoksia
8) lyhentää tiedon jalostusketjua
9) mahdollistaa takaisinkytkentä operatiivisiin järjestelmiin
10) saada tiedot omiin käsiin
11) yhdistää kaikki tiedot

Hankkeessa kirjattiin näiden lisäksi yleisiä tiedonhallinnan ratkaisuja koskevia tavoitteita:

1) helppo laajennettavuus
2) tietojen laatu hallinnassa
3) tietojen jäljitettävyys alkulähteille
4) kustannusten hallinta
5) riittävä suorituskyky latauksiin ja raporttien sekä kyselyiden vastausaikaan
6) tietojen kuvaukset riittävällä tasolla (metadata)
7) tarkoitukseen sopivat ohjelmistot (hinta, jatkuvuus, osaaminen, suorituskyky)
8) toimittajariippuvuuden (vendor-lock) vähentäminen
9) tulevaisuuden huomioiva, ei aseta rajoitteita ei-strukturoidun tiedon, kokeilevan ja reaaliaikaisen analytiikan suhteen

HUS Big Data Platform –hankkeelle määritellyt tavoitteet soveltuvat sellaisenaan tavoitteiksi myös tässä dokumentissa kuvattavalle HUS Tietoaltaalle.

Yleiskuva



Kuva 1. Yleiskuva Hadoop-pohjaisesta tietoallas-arkkitehtuurista ja ISAACUS rajapinnasta.

Johtava ajatus tietoaltaan teknisessä arkkitehtuurissa on hyödyntää modernien pilvipalveluiden tarjoamaa joustavuutta, kapasiteettia ja skaalautuvuutta, sekä avoimen lähdekoodin komponentteja tiedon analytiikkaan.

Modernit big data -järjestelmät rakentuvat Apache Hadoopin päälle. Tietoallas käyttää myös Apache Hadoopia. Perinteisesti Hadoopissa tieto on tallennettu klusterin tietokoneisiin ja hajautettu analytiikka ajetaan näillä tietokoneilla paikallisesti kyseisellä koneella sijaitsevaan dataan. Tämä lähestymistapa johtaa käytännössä siihen, että laskenta- ja tiedontallennuskapasiteetti ovat hyvin sidottuja toisiinsa, eikä niitä voida helposti erikseen skaalata.

Tietoallashankkessa käyttäjäryhmät eivät aseta jatkuvaa, raskasta laskentakapasiteettitarvetta tietoaltaaseen. Tutkimuskäyttö on hetkellistä ja voidaan etukäteen suunnitella siten että tarvittavat resurssit ovat saatavilla laskentaa tarvittaessa. Tämä puoltaa voimakkaasti tiedon tallennuksen ja laskentakapasiteetin erottamista toisistaan. Järjestely mahdollistaa paremman skaalautuvuuden useille tutkimusryhmille, koska jokainen ryhmä saa käyttöönsä oman taatun laskentakapasiteetin, jota voidaan ryhmäkohtaisesti tarvittaessa säätää.

Siirrettävyys

Tietoaltaan siirrettävyydessä on kaksi tasoa: Infrastruktuuri- ja ohjelmistotaso. Näiden kahden tason siirrettävyyttä voidaan tarkastella erillisinä komponentteina järjestelmän siirrettävyyttä arvioitaessa.

Ohjelmistotasolla tietoallas tulee seuraamaan avoimen lähdekoodin ratkaisuja (ks. Toimintaympäristö alla). Ohjelmistoratkaisut tiedon integraatiokerroksessa, ja tietoaltaan sisäisissä prosesseissa tuotetaan siten, että ne ovat ohjelmallisesti yhteensopivia yleisen, avoimen lähdekoodin Hadoop -ratkaisun kanssa. Tämä takaa niiden toimivuuden riippumatta toimittajasta, ja täten myös siirrettävyyden eri ympäristöihin.

Ohjelmistotason siirrettävyys mahdollistaa riippumattomuuden käyttäjäorganisaation infrastruktuuriratkaisuista. Tietoallasratkaisun ohjelmisto on asennettavissa sekä julkisiin pilvipalveluihin että perinteisiin konesaleihin perustuvaan infrastruktuuriin. Julkiset pilvipalvelut (Azure, AWS) ovat joustavan ja kustannustehokkaan käytön kannalta suositeltava infrastruktuuri tietoallasratkaisulle.

Tietoallasratkaisun käyttöönotto tapahtuu käyttöönottoprojektin kautta. Käyttöönottoprojektissa perustetaan tietoaltaalle tarvittava infrastruktuuri (tallennus, laskenta, tietoliikenne), perustetaan kehitys-, testi- ja tuotantoympäristöt sekä suoritetaan tietoaltaan modulaarisen ohjelmiston käyttöönotto tapauksesta riippuen tarvittavassa laajuudessa.

Vaatimukset ja käyttötapaukset

Toimintaympäristö

Avoin lähdekoodi ja Apache 2.0 -lisensoitavuus

Sitran Isaacus-hankekokonaisuus asettaa HUS Tietoaltaalle vaatimuksen lähdekoodin avoimuudesta ja Apache 2.0 lisensoitavuudesta. Vaatimuksella halutaan mahdollistaa HUS Tietoallas –ratkaisun hyödyntäminen myös muissa organisaatioissa HUS:n lisäksi kuten esimerkiksi Suomen muissa sairaanhoitopiireissä ja tulevaisuuden mahdollisissa laajemmissa terveydenhuollon organisaatioissa.

Vaatimuksen tulkitaan tarkoittavan, että

- Tietoallasratkaisun ohjelmiston lähdekoodi on avointa ja lisensoitavissa Apache 2.0 -lisenssillä
- Tietoallasratkaisussa hyödynnettävät valmistohjelmistot ovat lähdekoodiltaan avoimia ja Apache 2.0 -lisensoitavia
- Tietoallasratkaisun middleware-ohjelmistoalustat, joista Tietoallasratkaisun ohjelmisto on suoraan riippuvainen ovat lähdekoodiltaan avoimia ja Apache 2.0 -lisensoituja

Tietoallasratkaisun ohjelmistolla tarkoitetaan kaikkia ohjelmistopohjaisia osa-ratkaisuja sisältäen integraatio- ja muut sovellukset.

Avoimen lähdekoodin ja Apache 2.0 -lisensoitavuuden vaatimus ei koske middleware- eikä alemman tason ohjelmisto- ja muita teknisiä ratkaisuja

- joista Tietoallasratkaisun ohjelmisto on välillisesti riippuvainen (esim. käyttöjärjestelmät)
- joita Tietoallasratkaisun ohjelmisto hyödyntää standardoidun, ratkaisu-ja valmistajariippumattoman rajapinnan kautta (esim. tietokantapalvelimet)

Verkkoympäristö

HUS:n tietoverkko koostuu kotimaan datakeskuksesta, Azuren Pohjois-Euroopassa EU-alueella sijaitsevasta datakeskuksesta, näitä yhdistävästä HUS:lle dedikoidusta tietoliikenneyhteydestä (Azure ExpressRoute) sekä HUS:n toimipisteet kattavista paikallisista lähiverkoista. Tätä kokonaisuutta kutsutaan HUS:n sisäverkoksi. Pääsy julkisesta internetistä HUS:n sisäverkoon on estetty ilman asianmukaisia vahvoja tunnistautumis- ja muita tietoturvakäytäntöjä.

HUS Tietollas sijaitsee kokonaisuudessaan Azuren Pohjois-Euroopan tietokeskuksessa.

Tietoturva ja käyttöoikeudet

Tietoturvavaatimukset perustuvat HUS:n tietoturvapolitiikkaan ja HUS Tietoallas POC-hankkeelle asetettuihin erillisvaatimuksiin, sekä Euroopan ja Suomen henkilötietolainsäädäntöön. Näiden oletetaan pätevän myös HUS Tietoaltaan laajamittaiseen tuotantokäyttöön.

Yleisesti, tietoallas on terveystiedon toissijainen käyttäjä. Tämä tarkoittaa sitä, että tietoaltaan data ei ensisijaisesti sisällä uutta terveystietoa. Tietoaltaaseen kerätään dataa ensisijaisista järjestelmistä, joiden vastuulla on datan oikeellisuuden ja elinkaaren hallinta. Tietoaltaan data peilaa lähdejärjestelmien dataa, eikä siksi ole velvollinen toteuttamaan tiedon hallintaan liittyviä tehtäviä kuten uudessa Euroopan henkilötietosäännöstössä (General Data Protection Regulation, GDPR) [3] on määrätty.

Teknisesti tietoaltaan vaatimukset tulee täyttää yhteneväisesti GDPR:n kanssa. GDPR sisältää myös säännöstöä teknistä toteutusta ympäröiviin prosesseihin, joita tietoaltaan täytyy tukea.

1. Autentikointi

Autentikointi eli käyttäjien tunnistaminen täytyy olla luotettavaa erityisesti siinä tapauksessa, kun käsitellään pseudonymisoimattomia potilastietoja. Tietoaltaan tulee autentikoida käyttötapauksissa kuvatut suoria henkilö- ja järjestelmäkäyttäjät. Tietoallas ei autentikoi epäsuoria loppukäyttäjiä, kuten esimerkiksi tietoaltaaseen liitetyn analyysityökalun henkilökäyttäjiä vaan luottaa siihen, että työkalu on autentikoinut henkilökäyttäjänsä.

Koko tietoallasratkaisu rakentuu useasta pienemmästä tietoallasklusterista. Klustereiden rooleina ovat esimerkiksi tiedon automaattinen prosessointi, HUSin omille tutkijoille suunnattu analytiikka, ja yhteistyökumppaneille suunnattu analytiikka. Jokaiselle klusterille käyttäjähallinta voidaan tarvittaessa hoitaa eri tavalla. Pääsääntönä on kuitenkin, että käyttäjähallinta on liitetty HUSin Access Gateway (HAG) -palveluun.

2. Autorisointi

Autorisointi eli käyttöoikeuksien hallinta täytyy olla määriteltävissä ympäristössä varastoidun yksittäisen tietoaineiston tasolla kullekin yksittäiselle käyttäjälle tai käyttäjäryhmälle. Autorisoinnin yhteydessä käyttäjillä ja käyttäjäryhmillä tarkoitetaan tietoaltaan suoria henkilö- ja järjestelmäkäyttäjiä kuten edellä kohdassa ‘Autentikointi’ on kuvattu.

3. Auditointi

Auditoinnilla eli käytön valvonnalla on pystyttävä seuraamaan tietojen käyttöä yksittäisen tietoaineiston, käyttäjän ja käyttökerran tarkkuudella sekä tunnistamaan käyttöoikeuksien rikkomukset tai niiden yritykset. Auditoinnin yhteydessä käyttäjillä tarkoitetaan tietoaltaan suoria henkilö- ja järjestelmäkäyttäjiä kuten edellä kohdassa ‘Autentikointi’ on kuvattu.

Tutkijoilla ja tiedon hyödyntäjillä ei tule olla pääsyä pseudonymisoimattomaan dataan. Lisäksi lähdejärjestelmistä haettuun dataan hyödyntäjillä on vain lukuoikeus. Tämä vähentää auditoinnin tarvetta, koska tiedon lukuoikeus on määritelty tutkimusluvan puitteissa ja olemassa olevan tiedon muokkaus ei ole mahdollista.

4. Salaus

Salauksessa voidaan tarkastella erikseen tietojen salausta niitä siirrettäessä (data in motion) ja tallettaessa (data at rest). HUS:n tietoturvapolitiikan mukaan potilastiedot on salattava siirrettäessä niitä myös HUS:n verkon sisällä.

Tiedonsiirto tulee suojata TLS 1.2 ja myöhemmän standardin mukaisesti.

Tiedontallennus tulee toteuttaa salattuna AES256bit salauksella. Tämä koskee sekä Hadoop-osuutta että muita tilapäisiä tiedontallennuspaikkoja.

5. Pääsynhallinta

Tietoallas on suljettu ratkaisu, missä ulkopuolinen pääsy tietoaltaan tietoihin ja palveluihin on estetty sekä HUS:n verkon sisältä että sen ulkopuolelta, lukuun ottamatta erikseen määriteltyjä sallittuja tapoja käyttää tietoaltaan palveluita ja tietoa. Pääsy erikseen määritellyillä tavoilla tapahtuu ainoastaan luotettavan käyttäjän tunnistamisen kautta.

6. Asiakastietojen pseudonymisointi

Tietoaltaan hyödyntäjilleen tarjoamissa tietoaineistossa asiakastiedot tulee olla pseudonymisoitu siten, ettei asiakkaan henkilöllisyyttä ole mahdollista selvittää tietojen perusteella, mutta asiakkaaseen liittyvät tietoaineistot on mahdollista liittää toisiinsa asiakkaan anonyymisti yksilöivän avaimen perusteella. Poikkeustapauksissa asiakkaan henkilöllisyys tulee olla selvitettävissä anonyymin tunnisteavaimen perusteella.

Kahden tietoallastoteutuksen välillä tulee olla mahdollista yhdistää tietoa samaan henkilöön pseudonymisoinnista huolimatta. Tämä mahdollistaa analytiikan ja tietojen käsittelyn tietoallasrajojen yli. Pseudonymisointiratkaisu määritellään Isaacus-hankkeessa.

7. Jäljitettävyys

Tiedon prosessiointi pääaltaassa tapahtuu automatisoitujen ja versioitujen prosessien kautta. Tiedon kulkua pääaltaassa voidaan tarvittessa jäljittää seuraamalla automaatisoituja prosesseja ja niiden muodostamia ketjuja.

Tietointegraatiot

Tietolähdeintegraatiot

Tietoallasratkaisun keskeinen päämäärä on koota kaikki tieto samaan paikkaan ja mahdollistaa näin tietojen yhdistelyä vaativa tietojen hyödyntäminen eri tarkoituksiin. Tiedot kootaan Tietoaltaaseen ja ne pidetään siellä ajantasaisina tietolähdeintegraatioiden avulla. Integraatiot toteutetaan lähtökohtaisesti kaikkiin sisäisiin ja valikoituihin ulkoisiin tietolähteisiin priorisoidussa järjestyksessä. Sisäisen tietolähteen integrointi ei edellytä tunnistettua tarvetta tietolähteestä saatavilla olevaan tietoon. Tietolähdeintegraatiot tuovat tietolähteestä lähtökohtaisesti kaiken saatavilla olevan tiedon mahdollisimman alkuperäisessä muodossa mahdollisine laatu- ym. virheineen riippumatta siitä, onko kaikille tietolähteen tiedoille tunnistettu käyttötarpeita. Näiden periaatteiden kautta pyritään toisaalta saavuttamaan mahdollisuus kaikkien tietolähteiden tietojen yhdistelyyn ja analysointiin

välittömästi tarpeen ilmaantuessa sekä toisaalta mutkattomat ja kustannustehokkaat tietolähdeintegraatiot.

Integraatiot tiedon hyödyntäjien suuntaan

Tietoallasratkaisu kerää tietoa ja ajaa prosesseja, jotka jalostavat tietoa. Järjestelmän tulee myös kerätä metatietoa, joka kuvaa näitä tietoja mahdollistaen tietoaltaassa olevan tiedon ymmärtämisen ja löytämisen. Sekä kerätty ja jalostettu tieto että metatieto tulee olla saatavissa tietoaltaasta rajapintojen kautta tiedon hyödyntäjille. Lisäksi tietoallas toimittaa jalostettua tietoa hyödyntävien osapuolten tietovarastoihin.

Isaacus-integraatio

Isaacus-integraatio tarkoittaa Tietoaltaan valikoitujen tietojen sekä metatietojen tarjoamista kansallisen hyvinvoinnin palveluoperaattorin käyttöön. Palveluoperaattori on yksi Tietoaltaan keskeisimpiä hyödyntäjiä ja HUS:n Tietoallas vastaavasti yksi palveluoperaattorin keskeisiä tietolähteitä. Isaacus-kokonaisuus on kuvattu tarkemmin omassa dokumentissaan [1].

Isaacus-integraation vaatimukset ovat toistaiseksi avoimia seuraavassa lueteltujen asioiden osalta. Näiden vaatimusten odotetaan täsmentyvän tämän dokumentin kirjoitushetkellä käynnissä olevien Isaacus-esituotantohankkeiden tuloksena.

1. Mitä ovat palveluoperaattorin kannalta oleelliset tietoaineistot?.
2. Varastoiko palveluoperaattori fyysisesti tietoaineistoja?
3. Mikä on tietoaineistojen sisältämien henkilötietojen pseudonymoisointimalli palveluoperaattorin tasolla?
4. Miten metatietoja hallitaan palveluoperaattoritasolla?

Tiedonhallinta ja metatieto

Tietoaltaan tiedonhallinnan vaatimukset ovat johdettavissa tietoaltaan hyödyntäjiä ja hyödyntämistapoja kuvaavien käyttötapausten kautta (ks. Liite 2). Tarpeet tietoaltaan hyödyntämiselle ovat monipuoliset ja arkkitehtuurissa tulee varautua tarpeiden monipuolistumiselle tulevaisuudessa.

Metatiedonhallinnan perusvaatimus tietoaltaalle on tarjota tietoaltaaseen integroitujen tietolähteiden ja tietoaineistojen kuvaukset ja tekniset metatiedot tiedon hyödyntäjille helppokäyttöisen aineistoluettelon ja teknisen ohjelmistorajapinnan kautta. Tekninen metatieto tarkoittaa vähintäänkin tietoaineistojen kokoa, tietuemäärää, päivitystaajuutta ja viimeisintä päivitysajankohtaa sekä tietoaineistojen sisältämien tietoalkioiden tietotyyppejä.

Pääkäyttötapaukset

Käyttötapauksia voidaan kuvata eri tarkkuustasoilla ja eri näkökulmista. Pääkäyttötapaukset kuvaavat yhteenvedonomaisesti keskeiset tavat hyödyntää HUS Tietoallasta. Varsinaiset käyttötapaukset kuvaavat HUS Tietoaltaan hyödyntämistavat yksityiskohtaisemmin ja pääkäyttötapauksia tarkemmalla tasolla. Näiden lisäksi voidaan kuvata vielä teknisiä käyttötapauksia, joilla kuvataan HUS Tietoaltaan keskeiset toiminnot teknisestä näkökulmasta. Teknisiä käyttötapauksia käsitellään kappaleissa Tietoturva ja käyttöoikeudet sekä Lähdetiedot ja integraatiot.

HUS Tietoaltaan pääkäyttötapaukset ovat:

1. **Tutkimuksellinen käyttö**
   Tutkimuksellisen käytön tarkoituksena on mahdollistaa tutkimushankkeita tarjoamalla niille tapauskohtaisesti tietoaineistoja sekä mahdollisesti tiedonkäsittelytyökaluja ja laskentakapasiteettia. Tutkimushankkeet voivat olla kliinisiä tutkimuksia, HUS:n strategian ja operatiivisten toiminnan kehittämiseen liittyviä tutkimuksia sekä tuotekehitykseen liittyviä tutkimuksia. Tutkimushankkeita toteutetaan sekä HUS:n sisäisesti että ulkopuolisten yhteistyökumppaneiden kanssa.

2. **Toiminnanohjauksen raportoinnin tuki**
   Toiminnanohjauksen raportoinnin tarkoituksena on tuottaa tietoa HUS:n operatiiviseen toimintaan liittyvän päätöksenteon tueksi. Esimerkkinä toiminnanohjauksen raportoinnista voidaan mainita toiminnan kannalta keskeisten tunnuslukujen kuten potilaskuolleisuuden ja potilasreadmissioiden laskenta.

Varsinaiset käyttötapaukset on avattu *Liitteessä 2: Varsinaiset käyttötapaukset*.

Arkkitehtuuriratkaisu

Arkkitehtuuriratkaisu jakautuu viiteen pääalueeseen:

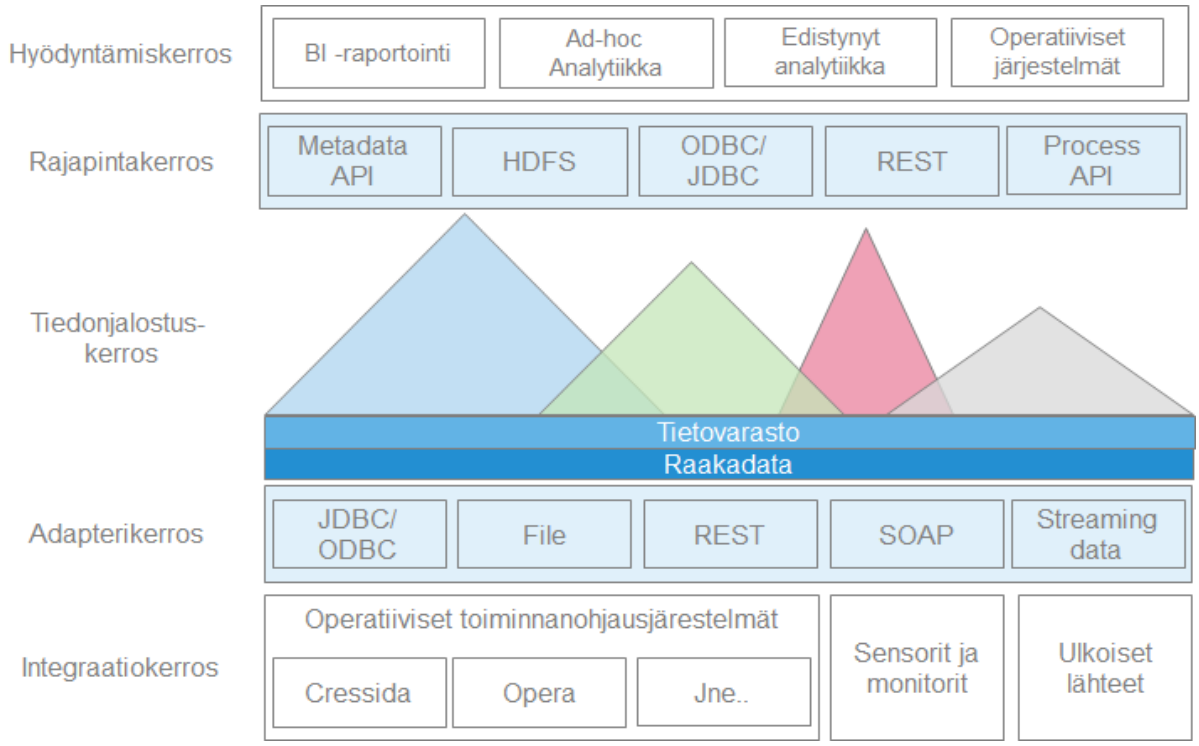
1. Tiedonhallinta
2. Infrastruktuuri
3. Ohjelmistoalusta
4. Integraatiot
5. ISAACUS integraatio

Tiedonhallinnan viitearkkitehtuuri

Seuraavassa kuvataan Data Lake –ratkaisuille tyypillinen viitearkkitehtuuri tiedonhallinnan näkökulmasta HUS:n toimintaympäristöön sovitettuna. Viitearkkitehtuuri voidaan toteuttaa monilla erilaisilla teknisillä ratkaisuilla. Seuraavassa kuvataan arkkitehtuurin toteutuksessa tavanomaisia ratkaisuja.

Yleiskuva

Viitearkkitehtuuri jäsennellään alla olevan kuvan mukaisesti eritasoisiin kerroksiin ensisijaisesti tiedon jalostusasteen perusteella.



Kuva 2. Tiedonhallinan viitearkkitehtuuri HUS Tietoallasprojektissa.

Tietoallas itsessään käsittää rajapinta-, tiedonjalostus- ja adapterikerrokset. Nämä on kuvassa korostettu sinisellä värillä.

Ylätasolla arkkitehtuuri jäsennellään hyödyntämis-, tietoallas- ja tietolähdekerroksiin, joista tietoallaskerros jäsennellään edelleen palvelurajapinta-, tiedonjalostus- ja lähtötietoadapterikerroksiin.

Seuraavassa käydään läpi kukin kerros sekä niiden sisältämät muut arkkitehtuurin elementit.

Hyödyntäminen

Arkkitehtuurin hyödyntämiskerrokseen kuuluvat tietoallasta hyödyntävät käyttäjät ja ulkopuoliset tekniset järjestelmät. Henkilökäyttäjien hyödyntämistavat jäsentyvät käyttötapauksiin, jotka on kuvattu tarkemmin Liitteessä 2. Henkilökäyttäjät hyödyntävät tietoallasta teknisten järjestelmien avulla. Lisäksi tekniset järjestelmät hyödyntävät tietoallasta automatisoitujen prosessien muodossa, jolloin hyödyntämisessä ei ole käyttäjiä suoranaisesti osallisena.

Tietoaltaan palvelurajapinta mahdollistaa tietoaltaan käytön erilaisilla työkaluilla ja järjestelmillä, jolloin kukin käyttäjäorganisaatio ja käyttäjä voi hyödyntää olemassa olevia tuttuja työkaluja.

Tietoaltaan hyödyntäjät voivat olla sekä tietoaltaan käyttäjäorganisaation (omistajan) sisäisiä että ulkoisia osapuolia. HUS EDW on tietoaltaan yksi merkittävimmistä sisäisistä hyödyntäjistä. Isaacus-palveluoperaattori on tietoaltaan keskeisin ulkoinen hyödyntäjä.

Tietoallas

Tietoallaskerros kuvaa ennen kaikkea tiedon jalostusketjun tietolähteistä tiedon hyödyntäjille. Näin ollen se kuvaa HUS Tietoaltaan sisäisen rakenteen loogisella tasolla tiedonhallinnan näkökulmasta. Tietoaltaan toimintaympäristö on monimutkainen. Erityyppisiä hyödyntämiskäyttötapauksia ja hyödyntäjiä kuten myöskin erityyppisiä tietolähteitä on paljon. Tämä asettaa tietoallaskerrokselle monipuolisia teknisiä ja toiminnallisia vaatimuksia. Vaatimuksiin vastaaminen edellyttää tietoallaskerroksen selkeää loogisen rakenteen määrittelyä ja määrittelyn noudattamista toteutuksessa.

Palvelurajapinta

Tietoaltaan palvelurajapintakerros sisältää tekniset palvelurajapinnat tiedon hyödyntäjien suuntaan. Sen tarkoituksena on määritellä ja kuvata rajallinen määrä tapauskohtaisesti valittuja ja tuotteistettuja palvelurajapintoja, joita tiedon hyödyntäjille tarjotaan.

Tiedonjalostus

Tiedonjalostuskerros jalostaa tietolähteistä hankittuja raakatietoja hyödyntämistä tukevaan muotoon. Tietolähteistä hankittuja tietoja yhdistellään yhtenäisiin, HUS:n käsitemalliin pohjautuviin tietorakenteisiin sekä tuottaa näistä jalostettuja tietoja laskentasääntöjä soveltamalla.

Lähtötietoadapterit

Lähtötietoadaptereilla tarkoitetaan tuotteistettuja tapoja toteuttaa ja operoida suuria määriä tietointegraatioita tuottavasti. Integraatioiden kautta tuodaan tietoaltaaseen erityyppistä tietoa erityyppisistä tietolähteistä. Adapteriratkaisuja kuvataan tarkemmin integraatioratkaisuja kuvaavan kappaleen alla.

Tietolähteet

Arkkitehtuurissa tietolähteet ovat tietoaltaan raakatiedon lähteitä. Tietolähteet voivat olla sisäisiä tai ulkoisia. Sisäisiä tietolähteitä ovat operatiiviset toiminnanohjausjärjestelmät, tietovarastot ja koneelliset tietolähteet kuten esimerkiksi palvelimet ja sensorit. Ulkoisia tietolähteitä ovat asiakkaiden, toimittajien ja muiden kumppanien operatiiviset toiminnanohjausjärjestelmät ja koneelliset tietolähteet sekä julkiset ja avoimet tietolähteet.

Tekninen infrastruktuuri

HUS Tietoallas - hankkeessa suunnitellaan järjestelmä, joka vastaa nykyajan tarpeita ja mahdollistaa joustavuuden ja laajennettavuuden tulevaisuudessa.

Päälinjaukset

Ratkaisun päälinjaukset ovat:

1. Tietoallas modularisoidaan hyödyntämällä Azuren PaaS -ratkaisuja. Ratkaisussa laskenta ja tiedon tallennus eriytetään toisistaan.
2. Azuren HDInisght Hadoop toimii Azure Data Lake Storagen (ADLS) kanssa, joita käytetään tiedon tallennusmediana.
3. Infrastruktuuri rakentuu päätietoaltaasta ja sivutietoaltaista. Päätietoltaan tarkoitus on toimia keskitettynä tiedon tallennuspaikkana, kun taas sivutietoaltaat toimivat alueina, jossa tutkimus ja kehitystyötä tehdään.
4. Sivutietoaltaita varten HDInisight -klustereita voidaan nostaa ja laskea tarpeen mukaan eri tutkimusryhmille ja muille toimijoille. Suurta laskentakapasiteettia vaativien yksittäisten integraatio- ja jalostustehtävien suorittamisessa voidaan hyödyntää myös tehtäväkohtaisia, kertakäyttöisiä klustereita.
5. Hyödynnetään pilvipalveluiden ohjelmallista infrastruktuurin ohjausta, jolloin klusterien hallinta voidaan suorittaa helposti skriptauksella tai muilla hallintavälineillä. Tällä säästetään kuluissa ja saadaan järjestelmän hallintaa vakioitua laadun parantamiseksi.
6. Rakennetaan vaatimusten mukainen käyttäjähallinta ja auditointi käyttäen Azuren komponentteja hyödyntäen Microsoftin tarjoamaa paikallista tukea.

Tutkijat, tutkimusryhmät tai muut sidosryhmät (yleisesti tutkijat) käyttävät tietoaltaan tiedon kanssa työskentelemiseen kahta erilaista tapaa:

- Tutkijoille perustetaan erillisen tietoallasinstanssi eli Tutkijan tietoallas (TT)
- Tutkojoille perustetaan vapaa virtuaalikoneklusteri, Tutkikan analytiikkaklkusteri (TA), joihin tuodaan valitut, tutkimukseen oleellisesti liittyvät tietoalkiot.

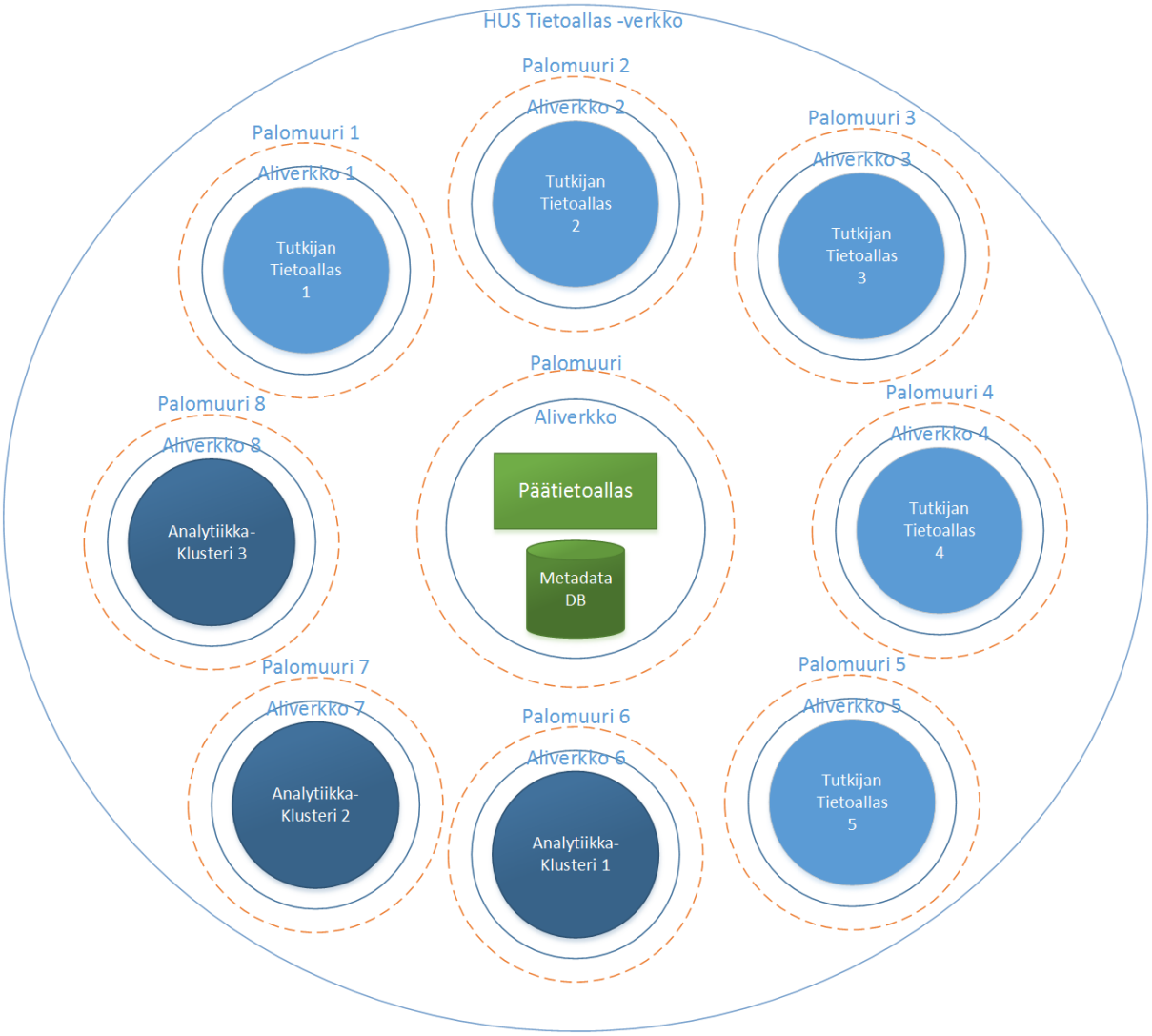
Tutkijan Hadoop-pohjaiseen tietoaltaaseen (TT) voidaan yhdistää tutkijoiden omia työkaluja ja jokainen TT on oma kokonaisuutensa. TT:hen tutkija voi tarvittaessa tuoda lisää tietoalkioita joko päätietoaltaasta tilaamalla, tai tuottamalla tietoalkioita itse.

Tutkijan analytiikkaklusteri (TA) on klusteri tietokoneita, joihin ei ole esiasennettu erityisiä ohjelmistoja. TA:n tarkoitus on tarjota erittäin joustava, mutta silti tietoallasympäristössä sijaitseva

ympäristö, jossa tutkijat voivat tehdä tutkimustyötään valitsemillaan työkaluilla asentamalla ne klusterin koneille.

Tekninen toteutus

Infrastruktuuri voidaan esittää kuvana missä keskellä on päätietollas, johon ulkopuoliset tietointegraatiot liitetään (Kuva 3). Tämä tietoaltaan osa sisältää kaiken tietoaltaaseen tallennetun tiedon ja sitä käytetään lähteenä kun tutkijan tietoallas-instansseja luodaan.



Kuva 3. Päätietoallas, metatiedon tallennustietokanta, ja tutkijan tietoaltaat ja analytiikkaklusterit suunnitellussa verkkoympäristössä.

HUS Tietoallas-projektissa hyödynnetään Azuren PaaS -komponentteja tuottamaan eristettyjä resursseja, tietoaltaita (TT) ja analytiikkaklustereita (TA). Eristettyjen resurssien lisäksi infrastruktuuri voi tarjota jaettuja palveluita. Keskeinen jaettu palvelu tiedon hallinnassa on metatiedon tallennus. Tämä mahdollistaa sen, että tietoaineistoon liitetty metatieto on suoraan käytettävissä kaikissa tietoaltaissa ilman erillistä lisätyötä. Näin myös tietoaltaiden käyttäjät saavat tietoonsa mitä

aineistoja on tarjolla, vaikka heillä ei olisikaan näihin aineistoihin lukuoikeutta. Metatietojen tallennus toteutetaan Apache Hiven Metastorella ja Azuren PaaS SQL tietokannalla.

Tietoallaskokonaisuuden resurssien eriytyksellä ja eristyksellä saavutetaan skaalautuvuutta ja tietoturvaa. Tätä tarkoitusta varten infrastruktuuri voidaan jakaa kolmeen pääalueeseen: Laskenta-, tiedontallennus- ja verkkoresursseihin.

Laskentaresurssien eristyksellä saavutetaan taattu laskentakapasiteetti eri tiloille. Lisäksi se mahdollistaa laskentaresurssien vapauttamisen kun niitä ei tarvita, tai vastaavasti lisäresurssien nopean lisäämisen jos käyttötapaus vaatii sitä. Lisäksi laskentaresurssien eristäminen tarjoaa mahdollisuuden “sandboxata” eri prosessit siten että niitä on helpompi tuottaa ja hallita.

Laskentaresurssien hallinta toteutetaan käyttäen Azure Resource Manager Templateita, joiden avulla erillisiä tietoallasklustereita on helppo hallita tarvittaessa vaikka kokonaan Azuren ulkopuolisesta hallintajärjestelmästä.

Tiedontallennusresurssien eristyksellä saavutetaan tietoturvaa, mutta myös mahdollisuus helposti jakaa tietoa usean eri tietoaltaan välillä. Eri tietoaltailla voi olla tietoon erilaisia oikeuksia jolloin samaa tietoa voidaan hyödyntää yhtä lailla testauksessa ja tuotekehityksessä.

Tiedontallennusresurssit toteutetaan käyttäen Azuren Blob Storagea, mutta myöhemmin tämä voidaan korvata Azure Data Lake Storagella sen valmistuttua tuotantoon EU-alueella.

Verkkoympäristön eristyksellä saavutetaan lisäturvaa palveluiden ja tiedon turvaamiseksi. Se sulkee tietoallaskokonaisuuden palomuurin taakse, mutta mahdollistaa HUSin ulkopuolisten toimijoiden liittämisen valittuun tietoaltaaseen esimerkiksi VPN-yhteyden avulla.

Päätietoallas

Päätietoaltaan tarkoituksena on toimia paikkana, johon tietoaltaaseen liitettyjen järjestelmien sisältämä raakatieto tuodaan ja käsitellään eteenpäin.

Päätietoaltaan vastuut ovat:

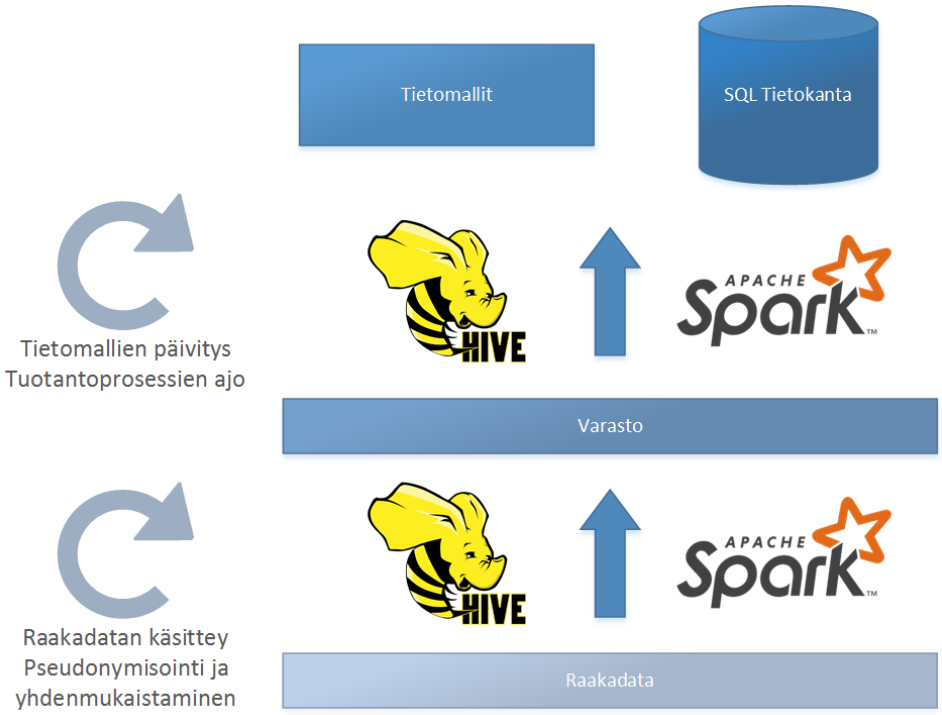
- Käsittelee integraatioiden tuottamaa dataa raakadatakerrosessa ja siirtää sen varastokerrokseen
- Huolehtii metadatan tallennuksesta ja päivityksestä
- Huolehtii tiedon yhdenmukaistamisesta ja pseudonymisoinnista
- Ajaa tiedonhallintaprosessit, jotka päivittävät tietoaltaassa olevaa tietoa ja rakentavat tietomalleja
- Ajaa tuotannossa olevia tiedonjalostusprosesseja ja integraatioita kohdejärjestelmiin

Päätietoaltaaseen ei ole pääsyä muilla kuin järjestelmän ylläpitäjillä. Kaikki tutkimus-, tuotekehitys- ja raportointikehitystyö tehdään tutkijan tietoaltaassa (ks. alla). Päätietoallas on jatkuvasti ajossa, mutta sitä voidaan skaalata ajon aikana ylös tai alas.

Ylläpitäjä voi muokata päätietoaltaan dataa. Tietoaltaassa täytyy pystyä tekemään perustason muutoksia dataan ja nämä toimet täytyy auditoida. Auditointi voi tapahtua joko tietoaltaan teknisillä järjestelmillä tai tietoaltaan ulkopuolisella prosessilla.

Päätietoallas toteutetaan erillisenä HDInsight -klusterina, jolla on kaksi erillistä tiedon tallennusaluetta:

1. Raakadatakerros - Sisältää tietolähteistä tulevan käsittelemättömän datan sellaisenaan
2. Varastokerros - Sisältää pseudonymisoidut ja esikäsitellyt tietoaineistot sekä jalostetut tietoaineistot.



Kuva 4. Päätietoaltaan toiminnallisuudet ja vastuut yhdistettynä tiedon siirtymiseen päätietoaltaassa.

Tutkijan tietoaltaat

Tutkijan tietoallas (TT) on oma tietoallas-instanssinsa, jonka sisältämä tieto on rajattu vain osaksi päätietoaltaan sisältämästä tiedosta. Valitut tiedot liitetään päätietoaltaasta tutkijan tietoaltaaseen sen luonnin yhteydessä joko linkittämällä ne suoraan vain lukuoikeudella päätietoaltaan varastotason (pseudonymisoituun ja yhdenmukaistettuun) tietoon, tai kopioimalla valitut tietoalkiot tai osia niistä. Tutkijan tietoaltaaseen voidaan myöhemmin tuoda mukaan lisää tietoalkioita. Tutkijan tietoaltaassa on myös oma tiedontallennuspaikka, johon tutkijat voivat tallentaa tutkimukseensa liittyvää tietoa. Tieto säilyy tallessa vaikka tietoallas poistettaisiin käytöstä,.

Tutkijan tietoaltaan tehtävät ja vastuut:

- Tutkimustyö - mahdollistaa tutkijoille joustava alusta, jolla voidaan tehdä tutkimustyötä sekä Hadoopiin kuuluvilla työkaluilla että ulkoisilla työkaluilla.
- Raporttikehitys - mahdollistaa perinteisen raportoinnin kehitystä käyttäen standardi-SQL -rajapintaa.

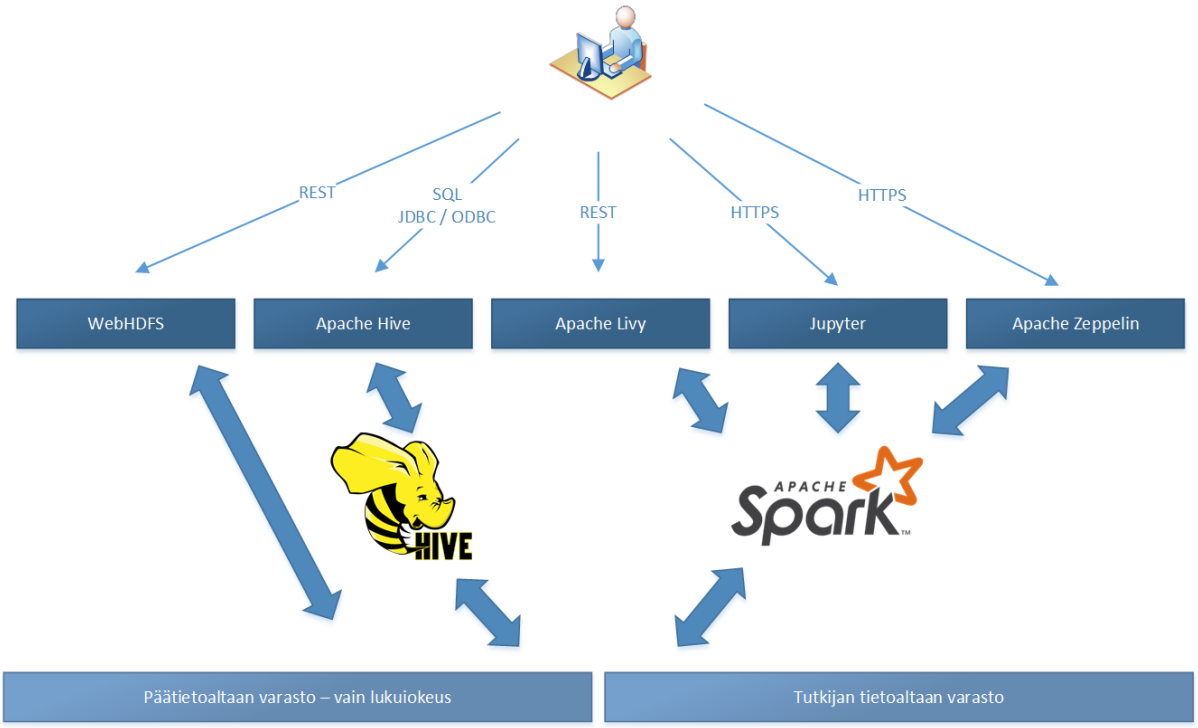
- Tuotteiden ja palveluiden kehitys - mahdollistaa tuote- ja palvelukehitys oikealla datalla tietoturvallisessa ja valvotussa ympäristössä.
- Yhteistyökumppanien työtila - mahdollistaa tutkimusyhteistyö

Tutkijan tietoaltaisiin on pääsy vain kyseisen tutkimusryhmän jäsenillä. Jokaiselle henkilölle on järjestelmässä oma käyttäjätunnus, jonka avulla käyttäjät tunnistetaan. Tutkijan tietoaltaassa on mahdollista rajoittaa eri henkilöiden pääsyä vain osaan siihen liitetystä datasta.

Tutkijan tietoaltaat ovat tilapäisiä ja niitä voidaan skaalata, nostaa ja laskea tarpeen mukaan. Tutkijan tietoallas ei siksi aja tuotantoprosesseja, ja jos tutkimuksen pohjalta halutaan tuottaa uusia tuotteita, raportteja tai palveluita, joiden halutaan olevan osa HUS tietoallaspalvelua, ne tuotetaan tuotteiksi erillisellä prosessilla ja liitetään päätietoaltaaseen (ks. yllä) ylläpitäjän toimesta.

Tutkija voi liittää TT:hen omia analytiikkatyökalujaan tiedon tutkimista varten. Ulkoiset työkalut voidaan liittää tietoaltaaseen käyttäen useita protokollia ja menetelmiä:

- JDBC/ODBC
- Apache Livy
- Apache Zeppelin
- WebHDFS



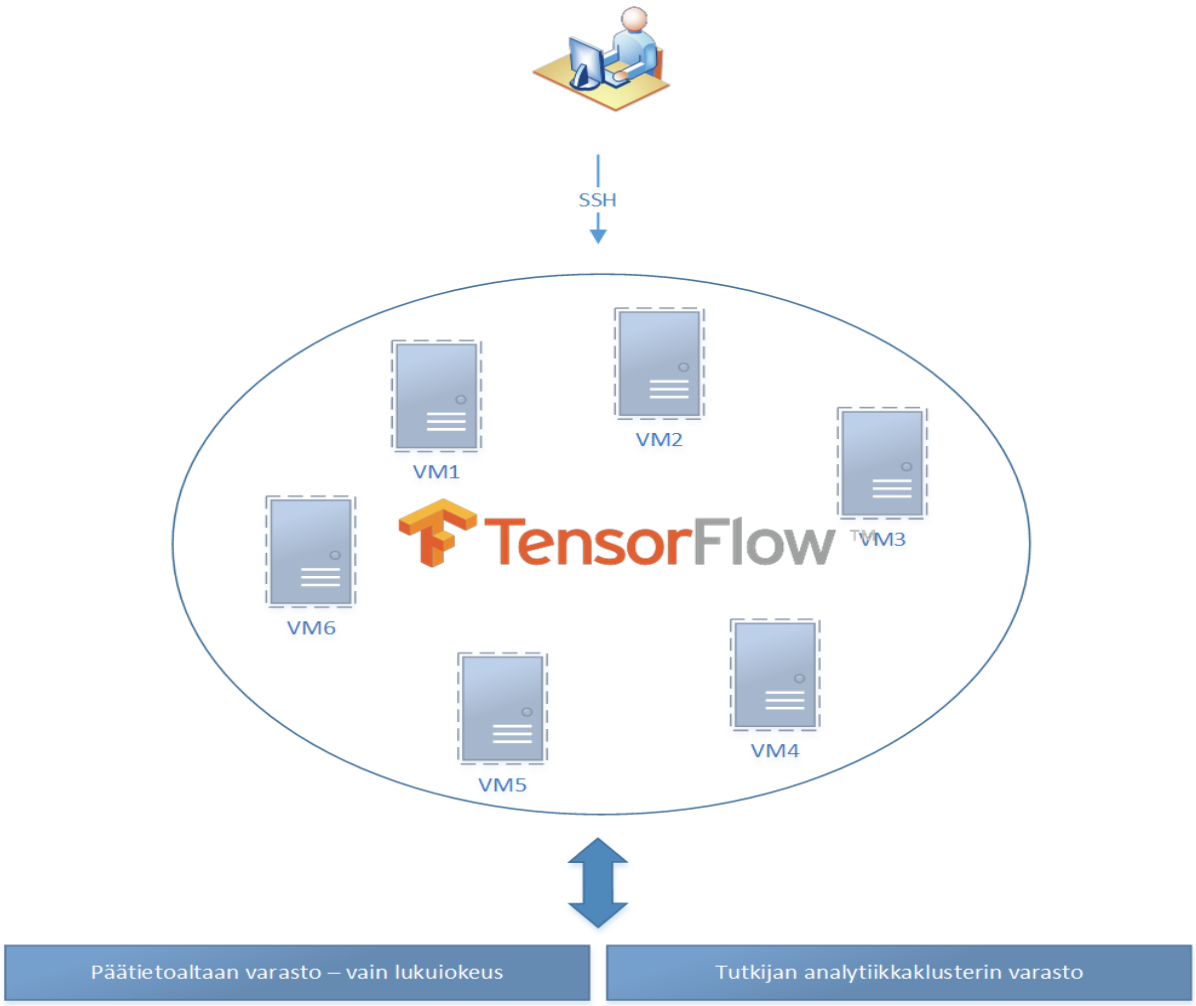
Kuva 5. Tutkijan tietoaltaan pääkomponentit ja rajapintatyypit

Tutkijan analytiikkaklusterit

Tutkijan analytiikkaklusterit ovat vapaamuotoisia virtuaalikoneklustereita, joiden tehtävänä on mahdollistaa hallitusti ei-Hadoop-pohjaisen analytiikan tekemistä tietoallasympäristössä.

Tutkijat saavat asentaa klusterin virtuaalikoneille vapaasti omia ohjelmistojaan ja heillä on pääsy päätietoaltaassa olevaan dataan. Analytiikkaklusterit eivät tarjoa valmiita ohjelmistoratkaisuita ja tiedon hakeminen päätietoaltaasta täytyy tehdä Azuren Blob storage ja Azure Data Lake Storage (myöhemmin) yhteensopivilla työkaluilla. Virtuaalikoneet liitetään HUSin Access Gateway-palveluun ja kirjautuminen virtuaalikoneille tapahtuu omilla käyttäjätunnuksilla. Analytiikkaklusterilla on myös oma tiedontallennuspaikkansa samalla tavalla kuin tutkijan tietoaltaassa. Tiedot pysyvät siellä vaikka klusteri suljettaisiin.

Esimerkki analytiikkaklusterista voi olla vaikka klusteri, jossa suoritetaan koneoppimista hyödyntäen tietoaltaan dataa ja käyttäen TensorFlow -ohjelmistoa (ks Kuva 6).



Kuva 6. Tutkijan analytiikkaklusteri. Esimerkkinä 6 virtuaalikoneen klusteri jossa asennettuna TensorFlow-ohjelmisto.

Ohjelmistoalusta

Hadoop-alusta

Hadoopin käyttäminen tietoaltaan ohjelmistoalustana vastaa Apache 2.0 -lisensointivaatimukseen, koska Hadoop sekä valtaosa eri Hadoop-jakeluihin sisältyvistä Hadoopin lisäohjelmistoista on lähdekoodiltaan avointa ja Apache 2.0 –mallilla lisensoitua. Vaatimus voidaan siis täyttää pysyttäytymällä Hadoop-ekosysteemin hyödyntämisessä avoimissa, yleisimmin käytetyissä, Apache 2.0 –lisensoiduissa ohjelmistoissa ja kehittämällä tietoaltaan sovellukset näiden kanssa yhteensopiviksi. Näitä ohjelmistoja ovat:

1. Apache HDFS
2. Apache Hive
3. Apache Spark
4. Apache Kafka
5. Hue (ei ole Apache-projekti, mutta Apache 2.0 -lisensoitu open source -ohjelmisto)

Näin tietoallasratkaisusta saadaan myös yhteensopiva eri Hadoop-jakeluiden kanssa PaaS-tyyppiset Hadoop-jakelut (AWS EMR, Azure HDInsight) mukaan lukien. Yleisimmät Hadoop-jakelut sisältävät pääosin edellä mainitut Apache-ohjelmistot. Yleisimpiä Hadoop-jakeluita ovat:

- Hortonworks (HDP)
- Cloudera (CDH)
- Amazon EMR
- Microsoft Azure HDInsights
- MapR
- IBM BigInsights

Tietojen tallennusratkaisu

Hadoop-yhteensopivat tiedostojärjestelmät ovat luonnollinen ratkaisu tietojen tallentamiseen hyödynnettäessä Hadoopia tietoaltaan keskeisenä ohjelmistoalustana. Hadoopin kanssa yhteensopivia tiedostojärjestelmiä on useita, mikä tarjoaa joustavuuden valita tapauskohtaisesti parhaiten soveltuva ratkaisu. Hadoopin oman hajautetun tiedostojärjestelmän HDFS:n (Hadoop Distributed File System) lisäksi yleisimpiä yhteensopivia tiedostojärjestelmiä ovat:

- Microsoft Azure Blob Storage ja Azure Data Lake Storage
- Amazon S3
- Google Cloud Storage
- MapR FileSystem

Tietojen tallennusratkaisuvaihtoehdoista seuraa muutamia merkittäviä etuja arkkitehtuurin näkökulmasta silloin, kun tallennusratkaisun suorituskyky ei ole tärkein valintakriteeri.

Tallennuskapasiteetti ja laskentakapasiteetti voidaan toteuttaa itsenäisinä ratkaisuina, mikä tuo merkittävää joustavuutta infrastruktuurin hallintaan sekä mahdollisuuden hyödyntää ylläpidollisesti vaivattomia elastisia PaaS-ratkaisuja kuten Azure Blob Storage, Amazon S3 ja Google Cloud Storage. Vaihtoehdot mahdollistavat myös tietoaltaan yhteensopivuuden yleisimpien julkisten pilvipalvelutarjoajien kanssa.

Integraatiot

Integraatioiden hallinta

Integraatioiden hallintamalli koostuu kolmesta tasosta:

- Orkestrointitaso
- Tehtävienhallintataso (aka. workflow-taso)
- Tehtävätaso

Orkestrointitaso

Orkestrointitaso koordinoi ja valvoo tiedonjalostusketjuja, niiden välisiä riippuvuuksia, tiedon laatua ja elinkaarta. Orkestrointitaso toteutetaan Apache Oozie coordinatoreilla Apache 2.0 lisensoitavassa ratkaisussa. Mikäli tapauskohtaisessa tietoallastoteutuksessa voidaan poiketa Apache 2.0 lisensoitavuusvaatimuksesta, tietoaltaan orkestointi voidaan myös toteuttaa muilla ratkaisuilla, esim. organisaatiossa yleisesti käytetyillä ETL- / integraatiotyökaluilla kuten esimerkiksi Talend, Pentaho, Informatica, Microsoft SSIS, tms. Kaupallisilla työkaluilla saavutetaan usein tuottavuusetuja Apache 2.0 -lisensoitaviin ratkaisuihin verrattuna monista syistä, mm. niiden korkeamman tuotteistusasteen johdosta ja koska voidaan hyödyntää organisaatiossa valmiina olevaa osaamista.

Tehtävienhallintataso eli workflow-taso

Workflow-tasolla hallitaan yksittäisistä tiedonsiirto- ja muokkaustehtävistä koottuja tiedonjalostusketjuja. Tehtävätaso toteutetaan Apache Oozie workflow'illa.

Tehtävätaso

Tehtävätasolla hallitaan yksittäisiä tiedonsiirto- eli integraatiotehtäviä sekä tiedonjalostustehtäviä. Sovellettavat ratkaisut riippuvat tehtävätyypistä.

Integraatiotehtävät

Tehtävätason integraatiotehtävien toteutus perustuu tietolähdeadaptereihin. Tietolähdeadapterit ovat nopeaa ja kustannustehokasta tietolähteiden integrointia tukevia tuotteistettuja integraatioratkaisuja. Tietolähdeadapterit jakautuvat teknisten integraatiotyyppien mukaisesti ylätasolla eräsiirto- ja sanomasiirtoadaptereihin sekä alemmalla tasolla integraatiotekniikan perusteella. Useimpien integraatiotekniikoihin kohdalla voidaan tehdä vielä lisäksi jako ns. pull- ja push-integraatiotyyppeihin.

Seuraavassa on lueteltu tyypillisiä data lake -ratkaisussa toteutettavia integraatioadaptereita. Päätökset HUS Tietoaltaaseen toteutettavista integraatioadaptereista tehdään tietolähteiden kartoituksen pohjalta.

1. Eräsiirtoadapterit

1.1. JDBC/PULL-adapteri relaatiotietokantaan perustuvien lähteiden tietojen lukemiseksi Tietoaltaaseen. Adapterin toteutus pohjautuu mikropalveluarkkitehtuuriin.
1.2. SOAP/PULL-adapteri SOAP-rajapinnan tarjoavien tietolähteiden tietojen lukemiseksi tietoaltaaseen. Adapterin toteutus pohjautuu mikropalveluarkkitehtuuriin.
1.3. REST/PULL-adapteri REST-rajapinnan tarjoavien tietolähteiden tietojen lukemiseksi tietoaltaaseen. Adapterin toteutus pohjautuu mikropalveluarkkitehtuuriin.
1.4. REST/PUSH-adapteri joka tarjoaa REST-rajapinnan tietolähteille tietojen kirjoittamiseksi tietoaltaaseen. Adapterin toteutus pohjautuu mikropalveluarkkitehtuuriin.
1.5. Fileshare/PULL- ja Fileshare/PUSH-adapterit
    1.5.1. Fileshare/PULL-adapteri tietolähteiden tietojen lukemiseksi tietolähteen tarjoamasta levyjaosta. Adapterin toteutus pohjautuu mikropalveluarkkitehtuuriin.
    1.5.2. Fileshare/PUSH-adapteri tietolähteiden kirjoittamien tietojen vastaanottamiseksi. Adapterin toteutus pohjautuu mikropalveluarkkitehtuuriin ja SFTP ohjelmistoon.

2. Streaming-adapteri
    2.1. REST/PUSH -adapteri streaming datan reaaliaikaiseksi vastaaottamiseksi reaaliaikaiseen tiedonvaihtoon kykenevistä tietolähteistä kuten esim. HUS:lla paljon käytetty Microsoft BizTalk Server. Adapterin Apache 2.0 -lisensoitavan ratkaisun toteutus perustuu Apache Kafka-ohjelmistoon.
    Kafka kykenee vastaanottamaan suuria tietomääriä nopeasti ja toimimaan puskurina tietoaltaan ja lähettäjän välillä. Lisäksi se integroituu helposti tietoallasteknologioiden kanssa ja sillä voidaan rakentaa helposti lähes-reaaliaikaisia tietovirtoja ja siihen littyvää analytiikkaa.

Muita yleiskäyttöisiä lähdetietoadaptereita toteutetaan tarpeen mukaan.

ISAACUS -integraatio

Isaacus-integraatioratkaisu kuvaa tavat tarjoata tietoaltaan tietoja sekä näiden metatietoja hyvinvoinnin palveluoperaattorin käyttöön.

Palveluoperaattorin käyttöön tarjottavien tietoaineistojen metatiedot julkaistaan THL:n esituotantohankkeen ja palveluoperaattorin toiminnallisten tarpeiden mukaisesti.

Toistaiseksi ratkaisuista koskien tietoaineistojen tarjoamista palveluoperaattorin käyttöön voidaan tehdä ainoastaan oletuksia, koska palveluoperaattorin tekniset toimintamalli on vielä avoinna.

Viitteet

1) ISAACUS - esituotantohankkeiden yhteinen projektisuunnitelma (Projekti 2018) v.0.2, Isaacus ps 0.2 3.6.16 luonnos.docx
2) Palveluväylä, kansallinen palveluarkkitehtuuri, http://vm.fi/palveluvayla
3) Euroopan uudet henkilötietoasetukset, General Data Protection Regulation, Regulation (EU) 2016/679

Tulevaisuus ja dokumentista rajatut seikat

Valvonta ja hallinta

Valvonta- ja hallintajärjestelmät ovat oleellisia osia minkä tahansa tietojärjestelmän tuotannossa. Tässä dokumentissa on tarkoituksella jätetty ottamatta kantaa jatkuvien palveluiden kysymykseen, mutta samalla kuitenkin ratkaisujen sunnittelussa on otettu huomioon valvonta ja hallinta.

Liitteet

Liite 1: Ei-toiminnalliset vaatimukset

Accessibility

Ei kuulu suunnitteluun.

Audit and control

- Käyttäjätunnistus AD/LDAPia vasten
- Kerberos -pohjainen käyttäjien tunnistus ja autorisointi
- Audit log

Availability (see service level agreement)

Tutkimuskäytössä olevan järjestelmän saavutettavuus on oltava yli 99%.

Backup

Tallennettu tieto tulee olla varmennettuna kolmeen kertaan.

Capacity, current and forecast

Kapasiteetti tulee mitoittaa siten, että tutkimushankkeissa käsiteltävien tietojen kokoluokka voi olla teratavuissa yhtä analyysiä varten.

Certification

Lopullisen järjestelmän tulee olla CE-hyväksytty.

Compliance

Järjestelmän tulee olle yhteensopiva ISAACUS -vaatimusten kanssa. Ks Open Source.

Configuration management

CE-hyväksyntä asettaa vaatimuksia järjestelmän dokumentoitavuudelle, järjestelmän konfiguraatiolle ja järjestelmän rakenteen seuraamiselle.

Dependency on other parties

Osa järjestelmän toiminnallisuudesta on sidottu ISAACUS-projektiin ja sen asettamiin vaatimuksiin.

Deployment

Asennussijaintina ja -alustana käytetään Microsoft Azure -pilvipalvelua. Järjestelmän tulee asentua nopeasti ja helposti.

Documentation

Järjestelmän dokumentaation tulee täyttää CE -vaatimus.

Disaster recovery

Tietoaltaista tulee ottaa varmuuskopiot soveltuvin osin.

Efficiency (resource consumption for given load)

Ei mukana suunnittelussa.

Effectiveness (resulting performance in relation to effort)

Ei mukana suunnittelussa.

Emotional factors (like fun or absorbing or has "Wow! Factor")

Ei mukana suunnittelussa.

Environmental protection

Ei mukana suunnittelussa.

Escrow

Ei mukana suunnittelussa.

Exploitability

Ei mukana suunnittelussa.

Extensibility (adding features, and carry-forward of customizations at next major version upgrade)

Analytiikka ja big data -järjestelmät kehittyvät ovat kehittyneet viime vuosien aikana paljon. Tämä asettaa järjestelmälle haasteita. Osa tekniikoista on edelleen jatkuvan kehityksen alla ja pitkän aikavälin ylläpidettävyyden näkökulmasta on oleellista taata järjestelmän laajennettavuus.

Failure management

Ei mukana suunnittelussa

Fault tolerance (e.g. Operational System Monitoring, Measuring, and Management)

Valvonta ja hallinta ei ole mukana suunnittelun tässä vaiheessa.

Legal and licensing issues or patent-infringement-avoidability

Ks. Open source

Interoperability

Yhteentoimivuus sekä lähdejärjestelmien että palvelun hyödyntämiskerroksen välillä tulee toteuttaa avoimia standardeja ja open source (Apache 2.0) tuotteita käyttäen.

Maintainability

Hadoopiin perustuvat ratkaisut ovat keskimääräistä monimutkaisempia ja vaikeammin ylläpidettäviä. Kokemusten ja osaamisen puute asettaa lisähaasteita ylläpidettävyydelle. Hyvän ylläpidettävyyden saavuttamiseksi suositaan mahdollisuuksien mukaan PaaS- ja SaaS-tason ohjelmistoratkaisuja itse asennettavien ja ylläpidettävien IaaS-tason ratkaisujen sijaan.

Modifiability

Ks. laajennettavuus

Network topology

Järjestelmän tulee olla yhteensopiva HUS tietoverkkorakenteen kanssa siten, että integraatiot HUSin järjestelmistä voidaan toteuttaa varmasti ja suojatusti.

Open source

Järjestelmän tulee perustua avoimen lähdekoodin lisensseihin (Apache 2.0). Vaatimuksena on että dataintegraatiot voidaan julkaista avoimena lähdekoodina. Järjestelmä itsessään tulee olla avoimien lähdekoodikirjastojen/APIen päälle rakennettu siten että järjestelmän kriittiset osat on mahdollista tuottaa täysin avoimen lähdekoodin tuotteilla.

Operability

Ei mukana suunnittelussa.

Performance / response time (performance engineering)

Ei mukana suunnittelussa.

Platform compatibility

Tietoaltaan tulee olla yhteensopiva Open Source Apache Hadoop -järjestelmien kanssa.

Price

Ei mukana suunnittelussa.

Privacy

Järjestelmään tallennettavat tiedot tulee pseudonymisoida.

Portability

Järjestelmä tulee voida siirtää toiseen palveluun siten että ohjelmistotasolla tietojärjestelmäintegraatiot ja järjestelmäprosessit toimivat uudessa ympäristössä ilman suurempia muutoksia.

Quality (e.g. faults discovered, faults delivered, fault removal efficacy)

Ei mukana suunnittelussa.

Recovery / recoverability (e.g. mean time to recovery - MTTR)

Ei mukana suunnittelussa.

Reliability (e.g. mean time between failures - MTBF, or availability)

Ei mukana suunnittelussa.

Reporting

Ei mukana suunnittelussa.

Resilience

Järjestelmän tulee pystyä selviytymään verkkokatkoksista ja palvelualustan uudelleenkäynnistyksistä siten, että palvelu palautuu normaaliin tilaan automaattisesti.

Resource constraints (processor speed, memory, disk space, network bandwidth, etc.)

Ei mukana suunnittelussa.

Response time

Ei mukana suunnittelussa.

Reusability

Järjestelmä tulee suunnitella komponenteista siten, että ne ovat uudelleenkäytettävissä sekä HUSin sisäisesti muissa älyprojekteissa että muiden sairaanhoitopiirien tietoallashankkeissa.

Robustness

Järjestelmän ei tule muuttua toimimattomaksi applikaatiotason virheiden sattuessa.

Safety or Factor of safety

Ei mukana suunnittelussa.

Scalability (horizontal, vertical)

Yksittäisten työskentelytilojen tulee olla skaalattavia sekä tiedon tallennuskapasiteetin että laskentatehon suhteen.

Security

Tietoallas tulee lähtökohtaisesti sulkea HUSin ulkopuoliselta pääsyltä täysin. Ulkopuolisia yhteyksiä tietoaltaaseen voidaan avata vain tarvittaessa. Tietoaltaan käyttäjätunnukset tulee sitoa HUSin yleiseen tunnuspolitiikkaan.

Tietoaltaan rakenteen tulee estää vihamielisen käyttäjän aiheuttamat vahingot jaetulle tiedolla tai toisten käyttäjien/tutkimusryhmien materiaaleille.

Software, tools, standards etc. Compatibility

Ei mukana suunnittelussa.

Stability

Tietoallas ja siihen liittyvät teknologiat ovat voimakkaan kehityksen ja muutoksen alla ja saattavat vaikuttaa järjesteläm pitkän aikavälin vakauteen (komponenttien muutoksiin).

Supportability

Ei mukana suunnittelussa.

Testability

Ei mukana suunnittelussa.

Usability by target user community

Kohderyhmän tulee osata tiettyjä teknlogioita, joille tietoallas on rakennettu. Minimivaatimuksen voitaneen pitää perus-SQL -tasoa. Tehokkaan käytön edellytyksenä on myös Apache Spark.

Liite 2: Varsinaiset käyttötapaukset

Käyttötapauksia voidaan tarkastella erilaisten roolien avulla. Tietoaltaan roolit ovat:

Tutkija, älypalvelu, raportointi, analytiikka, ylläpitäjä

Yleiset

1. Käyttäjä/palvelu kirjautuu tietoaltaaseen omilla tunnuksillaan. Tunnukset tulevat HUSin Active Directory –palvelusta.

Tutkija

Roolikuvaus: Tutkija jalostaa dataa uuteen muotoon ja analysoi sitä. Tieto voi olla jaettua tai tutkijan/tutkimusryhmän omaa dataa. Tutkija tekee raportteja, algoritmeja ja erilaisia löytöjä. Nämä tuotokset eivät itsessään ole vielä tuotantokelpoisia.

1. Tutkija saa oman tietoallas-instanssinsa tilaamalla sen ylläpitäjältä. Tietoallas mahdollistaa suurten datamäärien käsittelyn valikoitujen työvälineiden ja rajapintojen kautta. Tutkija voi tallentaa omaan tietoaltaaseensa dataa.
2. Tutkija käy dataa (jaettua tai omaansa) läpi eri työkaluilla. Pääasiallinen työkalu on Apache Hive, joka mahdollistaa SQL-rajapinnan käytön big dataan.
3. Tutkija haluaa pääsyn jaettuun dataan tietyn demografian tiettyyn dataan. Hän saa datasta kopion ja voi sen jälkeen vapaasti työstää dataa. Datan kopioinnista tutkijan tietoaltaaseen jää merkintä audit-lokiin.
4. Tutkijan käytössä oleva data on pseudonymisoitu, eikä siitä voi päätellä tutkimuskohteita yksilön tarkkuudella. Jokaisessa tietoaltaassa sama henkilö esiintyy eri pseudo-id:n kautta.
5. Jos lähdejärjestelmässä tieto muuttuu, muuttunut tieto siirtyy tietoaltaaseen viivellaä change capture –järjestelmän kautta.
6. Tutkija voi olla myös ulkopuolinen toimija (esim. yritys tai julkinen toimija), jolle annetaan pääsy rajattuun tietomäärään. Ulkopuolinen toimija voi tällöin suorittaa analytiikkaa vapaasti yhdistäen omaa tietoaan tietoaltaassa jo olevaan tietoon.

Älypalvelu

Roolikuvaus: Älypalvelu on palvelu, joka käyttää tietoallasta itsessään tai tietoaltaan pohjalta tuotettuja palveluja. Älypalvelu tarvitsee aina palveluarkkitehtuurin toimiakseen.

1. Älypalvelu tekee virtaavalle (streaming) datalle analyysiä ja reagoi tiettyihin muutoksiin tietovirrassa. Tällaisia käyttötapauksia ovat esim. keskosten sepsis.
2. Älypalvelu tarjoaa APIn asiakkailla ja suorittaa ennalta laskettua mallia vastaan tuloslaskennan ja palauttaa ”reaaliaikaisesti” tuloksen kutsujalle. Palveluesimerkkinä analyysi lääkkeiden yhteiskäyttövaikutuksesta.
3. Älypalvelu suorittaa suuremman laskennan tai tiedon käsittelyn tietoaltaassa. Palvelu voi joko palauttaa vastauksen tai käynnistää muita prosesseja, jotka tuottavat tuloksen kohdejärjestelmään.

Raportointi ja raportin kehitys

Roolikuvaus: Raportointi-rooli viittaa perinteiseen BI-raportointiin ja EDW liitäntään, ja muihin vastaaviin järjestelmiin. Nämä järjestelmät vaativat nopean SQL-pohjaisen rajapinnan. Raportin

kehittäjän voivat kehityksen yhteydessä haluta suorittaa ad hoc –tyylisiä kyselyitä tietoaltaan tietoon.

1. Ulkopuolinen raportointijärjestelmä haluaa raporttiinsa tiedon analyysin pohjalta tehdyistä tiedoista. Analyysi ajetaan aikataulun mukaan vaikka joka yö ja raportointijärjestelmä lukee analyysin tuloksen SQL-palvelimelta.
2. Raportin kehittäjä haluaa raportin kehityksen yhteydessä suorittaa satunnaisia SQL-kyselyitä tietoaltaaseen.

Analytiikka

Roolikuvaus: Analytiikka tarkoittaa tässä yhteydessä tapauskohtaista ad-hoc-tarpeista lähtevää tiedon analysointia sekä toimenpidesuositusten, vastausten sekä muiden johtopäätösten tuottamista analyysin pohjalta. Analytiikka voi olla reaktiivista ja proaktiivista.

1. HUS:n yksiköt tarvitsevat faktapohjaisen päätöksenteon tueksi tietoa, jota raportointi ei tarjoa ja jota heillä muutenkaan käytettävissään. Yksikkö määrittelee tietotarpeensa yhteistyössa analyytikon kanssa, joka hankkii tarvittavan tiedon Tietoaltaaseen koottuja tietoja yhdistelemällä ja analysoimalla.
2. Analyytikko muodostaa hypoteeseja HUS:n toiminnan kannalta keskeisistä kehittämismahdollisuuksista, verifoi hypoteeseja Tietoaltaaseen koottuja tietoja yhdistelemällä ja analysoimalla sekä muodostaa analyysin tulosten pohjalta yhteenvetoja ja toimenpidesuosituksia HUS:n yksiköille.

Ylläpitäjä

Roolikuvaus: Ylläpitäjä hoitaa tietoaltaaseen liittyvä hallinnallisia toimenpiteitä

1. Tutkija tarvitsee uuden tutkijan tietoaltaan. Ylläpitäjä perustaa tietoaltaan ja sopii tutkijan kanssa työtilan resurssitarpeesta.
2. Tutkija tarvitsee tietoaltaan yleista dataa. Ylläpitäjä kopioi data tutkijan tietoaltaaseen ja huolehtii että se näkyy oikealla tavalla tutkijan työtilassa.
3. Tutkija lopettaa tutkimuksensa. Ylläpitäjä poistaa työtilan pois käytöstä ja siivoaa siihen liittyneet resurssit.