Kohti kansallista sote-tietojen saatavuuspalvelua

Sosiaali- ja terveydenhuollon aineistojen saatavuusarkkitehtuurin skenaario ja valmisratkaisujen testaus

Väliraportti 2

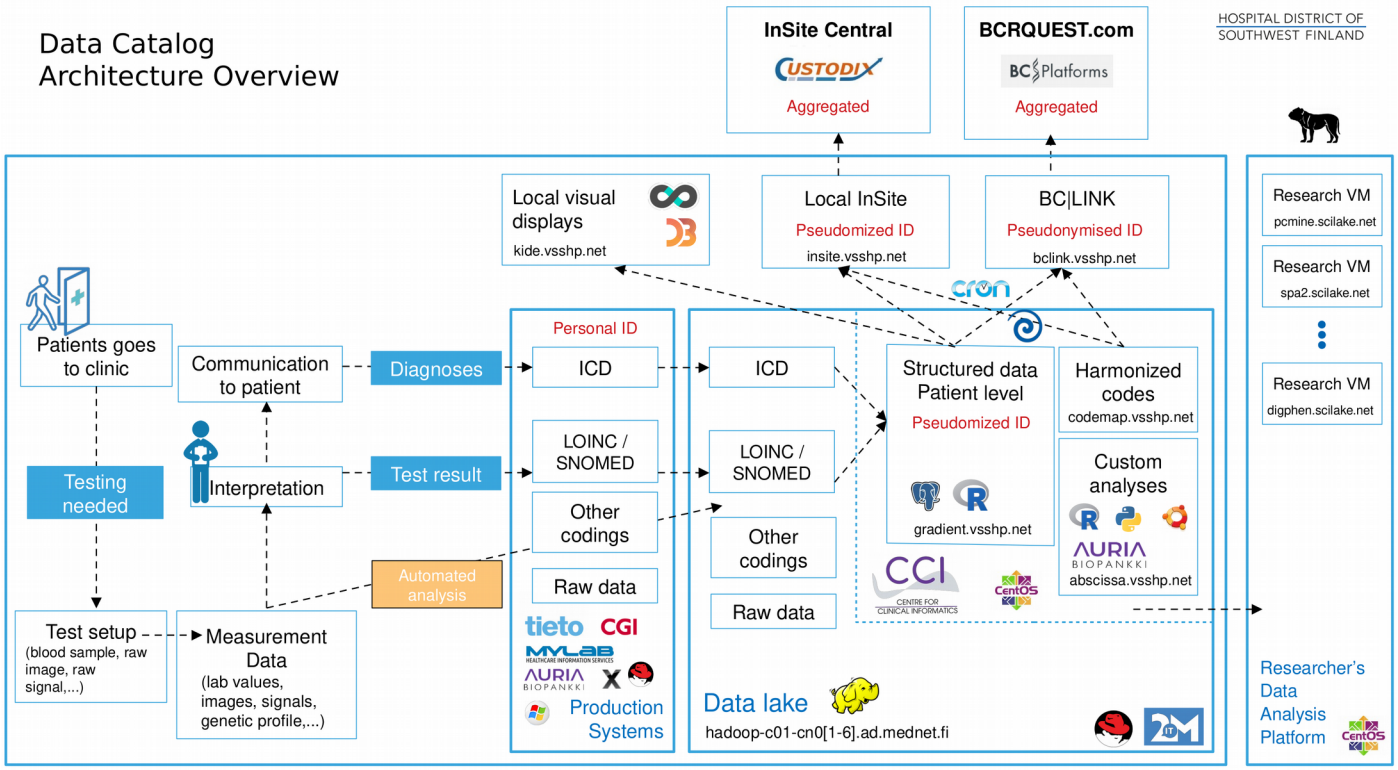
Versiohistoria:

Versio	Päivämäärä	Valmiusaste (vedos/ehdotus/päivitys)	Kirjoittaja(t)	Huomautukset
0.9	2018-04-04	Vedos	Arho Virkki	Ensimmäinen versio
1.0	2018-04-05	Ehdotus	Arho Virkki	Täydennetty versio

Sisällys

Arkkitehtuuriskenaario

1.1 Kokonaisuuden kuvaaminen



Kuva 1. Tieto-, järjestelmä- ja teknologia-arkkitehtuurin kuvaus saatavuuspalvelun käyttäjän näkökulmasta (VSSHP).

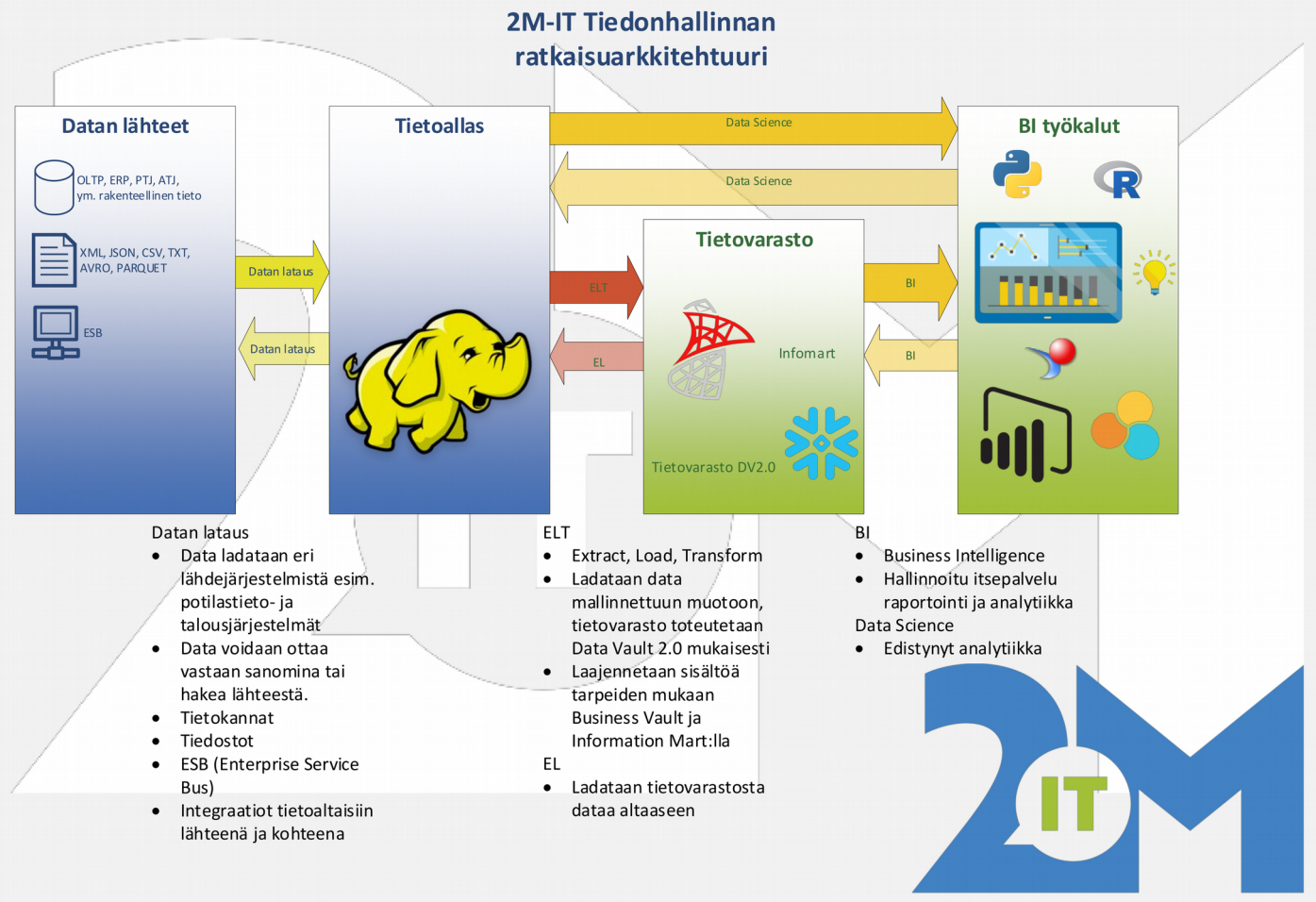
1. Tieto kertyy operatiivisiin järjestelmiin, josta se ladataan ajastetusti M2-IT:n ylläpitämään Hadoop-pohjaiseen raakatietoaltaaseen (Cloudera Hadoop ja Red Hat Enterprise Linux).

2. Seuraavaksi VSSHP:n Kliininen tietopalvelu (Centre for Clinical Informatics; CCI) yhteismitallistaa datan rakenteiksi, joista ladataan kaikkien saatavuuspalveluiden ja käyttöliittymien raakatieto (kuvassa paikallinen Custodix InSite, ja BC|Link). Koodistojen harmonisointi kannattaa tehdä keskitetysti: Tätä varten on erillinen palvelin (codemap.vsshp.net), jonne jokaisella koodistojen käsittelijällä on luku- ja kirjoitusoikeus.

3. Paikalliset käyttöliittymät (Kide; Apache Superset) ja analyysiympäristö (Kliininen tietopalvelu, talouspalvelut ja Auria Biopankki; abscissa.vsshp.net; Ubuntu 16.04, R ja Python) käyttävät samaa yhteismitallistettua dataa.

4. Saatavuutta on mahdollista tutkia tätä kirjoittaessa jo sisäisen InSite-palvelun kautta, jonne tiedot päivittyvät kerran päivässä. Käyttöliittymään pääsee sairaalan sisäverkosta ja siihen kirjaudutaan erikseen. Muun muassa arviointiylilääkärillä on valtuudet lisätä palveluun käyttäjiä.

5. Tutkimusluvalla suoritettavaan tutkimukseen on mahdollista perustaa tutkijan allas (Researher's Data Analysis Platform; DAP). Data luovutetaan myös tähän ympäristöön samasta yhteismitallisesta joukosta, jota käytetään saatavuuspalveluiden pohjana. Tutkimusympäristöön voi kirjautua RSA-avaimen ja salasanan yhdistelmällä ja se avataan tietosuojasitoumusta vastaan.



Kuva 2. Ylätason kuvaus lähdejärjestelmien, tietoaltaan, erikoistuneiden analytiikkatietokantojen ja käyttöliittymien suhteista (2M-IT Oy).

Tiedonhallinta koostuu karkeasti neljästä osasta:

1. Operatiiviset järjestelmät, jotka ovat tiedon ensisijaisia lähteitä.
2. Tietoallas, johon tieto ladataan ajastetusti (tyypillisesti kerran vuorokaudessa), ja jota voidaan teknisesti käyttää suoraan tiedon mallintamiseen ja analyysiin.
3. Tietovarastot, jotka ovat erikoistuneet raportointiin (SQL Server) tai data-analyysiin (PostgreSQL + upotettu R-kieli), ja jotka on optimoitu tarkoitukseensa,
4. Erilaiset selainpohjaiset analyysi- ja raportointikäyttöliittymät, kuten PowerBI, SAP BO, RStudio, Jypyter ja Apache Superset.

Testauskohteiden valinta ja toteutus

2.1.4 Testaus, paikallinen ja kansainvälinen, ja jatkosuunnitelma

- Käyttöliittymiä tullaan testaamaan ristiin: Custodixin, BC Platformsin sekä suoraan altaaseen suoritetun tietohaun pitäisi tuotta sama tulos.
- Halukkaat teollisuuskumppanit on etsitty yhdessä BC Platformsin kanssa: Pfizer, Biogen ja Amgen.

2.2.1 Tietolatausten automatisointi

Aikaisemmin toteutettu kertalataus kuvan 1 mukaisesti on nyt muutettu toimimaan päivittyvästi kerran vuorokaudessa.

2.2.2 Koodistojen siltaus kansainvälisiin koodeihin

Tämä kannattaa tehdä kansallisesti ja THL-vetoisesti, yhdessä Suomen Biopankkiosuuskunnan Saatavuuspalvelu-hankkeen kanssa. THL:llä oli tiedotustilaisuus aiheesta 4.4.2018.

2.2.4 Testataan terminologiapalveluna THL käytössä olevaa Code Server – tuotteen VSSHP installaatiota ja tunnistetaan THL koodistojen täydennystarpeet

Tämä kannattaa tehdä kansallisesti ja THL-vetoisesti, yhdessä Suomen Biopankkiosuuskunnan Saatavuuspalvelu-hankkeen kanssa. Meillä on vahva epäilys, että nykyinen Code Server -teknologia on vanhentunutta, mutta asia tarkistetaan.

2.2.7 Kuvataan ja demonstroidaan BC Platforms ratkaisun arkkitehtuuri ja sen erot toiseen ratkaisuun

Järjestelmät ovat pitkälti samanlaiset, kuten Kuva 1. osoittaa. Tärkein ero näiden kahden järjestelmän välillä on, että

1. Custodixin järjestelmä muodostaa tilaston paikallisesti ja yhdistää tiedot myöhemmin
2. BC Platformsin järjestelmä muodostaa tilastot kyselyiden perusteella

Tästä johtuen BC Platformsin järjestelmässä

- voidaan tehdä mutkikkaampia (ja siten rikkaampia) kyselyitä
- tarvitaan enemmän laskentatehoa kuin Custodixin InSitessa
- on vaikeampi todistaa formaalisti, että se täyttää anonymisoinnin vaatimukset (peräkkäisten kyselyiden ja siksi tarkentuvan tiedon ominaisuudesta johtuen).

3 Kansallisen toteutuksen tiekartta ja riskit

Tämä kannattaa tehdä kansallisesti yhdessä Suomen Biopankkiosuuskunnan Saatavuuspalvelu-hankkeen kanssa. Tällä hetkellä ehdotamme, että

- THL vastaa terminolgiapalvelusta.
- Analyysi kansallisista haasteista tehdään, kun kansallista yhteistyötä on ensin kokeiltu Biopankkiosuuskunnan vastaavassa STM-hankkeessa.
- Geentiedon ja muun massadatan yhdistämisestä tiedetään jo paljon, mutta tätä varten haluamme haastatella vielä CSC:n asiantuntijoita.

3.2 Välttämättömät sairaalakohtaiset tekniset edellytykset

Kohdan 1.1. arkkitehtuuriskenaario sekä Kuva 1 ja Kuva 2 esittävät välttämättömät sairaalakohtaiset tekniset edellytykset. Tutkijan allas ei ole välttämätön, mutta muut tiedon yhteismitallistamisen vaiheet tarvitaan.