**lIsa 2 Põllumajanduse Suurandmete süsteemi  
KONTSEPTSIOON**

Põllumajanduse Suurandmete projekt



Sisukord

[1. Mõisted ja lühendid 4](#_Toc17990052)

[2. Ülevaade 4](#_Toc17990053)

[3. Millistele parameetritele peab suurandmete süsteem vastama 5](#_Toc17990054)

[1.1 Andmehõive ja töötluse süsteem 6](#_Toc17990055)

[1.1.1 Andmete standardne vorm 7](#_Toc17990056)

[1.2 Paralleelprotsesside jooksutamine taustal 12](#_Toc17990057)

[1.3 Analüütikakeskkond 13](#_Toc17990058)

[1.4 Teenuste veebikeskkond 15](#_Toc17990059)

[1.5 Suurandmete süsteemi haldamise keskkond 16](#_Toc17990060)

[1.6 Suurandmete süsteemi nõuete loetelu 18](#_Toc17990061)

[2 Suurandmete süsteemi andmed 32](#_Toc17990062)

[2.1 Suurandmete süsteemi poolt kasutatavad andmekogud 32](#_Toc17990063)

[3 Arhitektuurivariandid 35](#_Toc17990064)

[3.1 Kogumine vs ühendumine 35](#_Toc17990065)

[3.2 Pilv vs lokaalne paigaldus 40](#_Toc17990066)

[3.3 Toega vs toeta baastarkvara 41](#_Toc17990067)

[3.4 Suletud lähtekood vs avatud lähtekood 42](#_Toc17990068)

[3.5 Andmevahetus andmekogudega – veebiteenused, andmebaasiühendused, failid 42](#_Toc17990069)

[3.6 Andmevahetus põllumajandusmasinatega 42](#_Toc17990070)

[4 Tehnilise arhitektuuri variandid 48](#_Toc17990071)

[4.1 Andmebaaside jõudlusest 48](#_Toc17990072)

[4.2 Lahenduse horisontaalne skaleeruvus 49](#_Toc17990073)

[4.3 Arhitektuuri variant nr 1 49](#_Toc17990074)

[4.4 Arhitektuuri variant nr 2 50](#_Toc17990075)

[4.5 Arhitektuuri variant nr 3 51](#_Toc17990076)

[4.6 Arhitektuuri valimine 53](#_Toc17990077)

[5 Suurandmete süsteemi standardite kirjeldus 57](#_Toc17990078)

[5.1 Füüsilisele andmemudelile esitatavad nõuded 57](#_Toc17990079)

[5.2 Mõõdikute kirjeldusele esitatavad nõuded 58](#_Toc17990080)

[5.3 Dimensioonide kirjeldusele esitatavad nõuded 59](#_Toc17990081)

[5.4 Klassifikaatoritele esitatava nõuded 59](#_Toc17990082)

[5.4.1 Tabel: KLASSIFIKAATOR 60](#_Toc17990083)

[5.4.2 Tabel: KL\_ELEMENT 61](#_Toc17990084)

[5.4.3 Tabel: KL\_VERSIOON 62](#_Toc17990085)

[5.4.4 Tabel: SEOS 63](#_Toc17990086)

[5.5 Riiklikud nõuded metaandmete kirjeldamise osas 63](#_Toc17990087)

[5.6 Teenuste kirjeldustele esitatavad nõuded 63](#_Toc17990088)

[5.7 Täiendavad soovitused kirjelduste koostamiseks 64](#_Toc17990089)

[5.8 Veebiteenuste standardid 64](#_Toc17990090)

[5.9 Andmetüüpide kasutamine 66](#_Toc17990091)

[5.10 Andmekvaliteedi standard 66](#_Toc17990092)

[5.11 Andmekogu kvaliteedi hindamine 68](#_Toc17990093)

[6 Suurandmete süsteemi teenused 69](#_Toc17990094)

[6.1 Teenuste valik 69](#_Toc17990095)

[6.2 Teenuste majandusanalüüs 70](#_Toc17990096)

[7 Suurandmete süsteemi loomise teekaart 76](#_Toc17990097)

[8 Viited 79](#_Toc17990098)

# Mõisted ja lühendid

|  |  |
| --- | --- |
| Mõiste nimi | Selgitus |
| ADS | Eesti aadressandmete süsteem |
| Andmekogu | Ametlikult RIHA-s registreeritud või ametlikult registreerimata andmestik |
| MeM | Maaeluministeerium |
| PMA | Põllumajandusamet |
| PMAIS | Põllumajandusameti infosüsteem |
| PMK | Põllumajandusuuringute Keskus |
| PRIA | Põllumajanduse Registrite ja Informatsiooni Amet |
| RIA | Riigi Infosüsteemide Amet |
| RIHA | Riigi infosüsteemi haldussüsteem. Tegemist on metaandmete süsteemiga, kus peavad olema kirjeldatud infosüsteemid, mis kuuluvad riigiinfosüsteemi koosseisu. |

**Tabel 1. Mõisted ja lühendid**

# Ülevaade

Eestis on alustatud mitmes valdkonnas suurandmete süsteemide loomist. Hetkel puuduvad aga tavad ja standardid, kuidas selliseid süsteeme tuleks luua. Samuti ei ole üheselt kokku lepitud, mis on suurandmed. Selle tõttu on käesolevas projektis otsitud koostööd erinevate haldusalade vahel, et leida võimalikult laiahaardeline lahendus ja üritada rakendada kogu teemat puudutavat teadmust, mis seni riigis tekkinud on.

Suurandmeteks on käesolevas projektis peetud andmeid, mis vastavad järgmistele tunnustele:

1. andmete lisandumise kiirus on suur;
2. andmete keerukus on suur;
3. andmete maht on suur;
4. andmete sisu ei ole eriti täpselt teada või andmed on struktureerimata.

Selliseid andmeid, mis vastaksid kõikidele nendele kriteeriumitele, analüüsi käigus leida ei õnnestunud. Hindamiskriteeriumid on kindlasti ka veidi pehmed ja subjektiivselt hinnatavad, kuid siiski muu maailma mõttes on Eestis toodetavad põllumajanduse andmed veel väikesed.

Küll aga õnnestus leida suurandmete tulevased allikad, milleks on peamiselt põllumajandusmasinate andmed ehk põllutööde andmed, maapinna ja mulla omaduste andmed detailsel tasemel ning satelliidipiltide põhjal arvutatavad indeksid, kui need esitada detailsusega iga m2 maapinna kohta. Selliste andmeid täna veel üleriigiliselt ei kasutata ja pole ka välja töötatud süsteeme, mis selliseid andmeid registreerivad ja kasutavad. Esimesena paistab tekkivat põllutööde andmete süstemaatiline registreerimine elektroonilisel kujul ning satelliidipildi pealt indeksite arvutamine, mis iseloomustavad taimekasvatuses toimuvat.

Käesoleva Suurandmete süsteemi kontseptsiooni aluseks on põllumajandusvaldkonnaga seotud andmekogude andmeanalüüs, õiguslik analüüs ja majandusanalüüs sh Suurandmete süsteemi võimalike teenuste analüüs.

Suurandmete süsteemi andmed jagunevad järgmistesse kategooriatesse:

1. Reaalandmed objektide ja subjektide kohta. Suurandmete süsteem sisaldab osasid reaalandmeid, mis liidestatud andmekogud ei sisalda ja mille jaoks muu andmekogu puudub. Samuti võidakse suuremate analüüside tegemise ajaks kopeerida Suurandmete süsteemi erinevate andmekogude andmeid, et analüüs oleks võimalik teostada reaalajas ning analüüsi oleks võimalik korrata täiendatud algoritmidega.
2. Reaalandmete ja teenuste kirjeldused ehk metaandmed.
3. Isikute (süsteemide kasutajate) ja nende õiguste ning volituste info.

Suurandmete süsteem ei hakka püsivalt sisaldama teiste andmekogude andmeid sh ruumiinfot, mille jaoks on Maa-amet arendanud välja tugeva infrastruktuuri ja üldkasutatavad komponendid, millele ei ole mõtet hakata looma konkureerivat lahendust. See kehtib ka Maa-ameti poolt loodud ESTHubi satelliidipiltide töötlemise keskkonna kohta, mille võimekust on plaanis hakata tarbima erinevate indeksite väljatöötamisel ja arvutamisel.

# Millistele parameetritele peab suurandmete süsteem vastama

Suurandmete süsteem jaguneb järgmisteks osadeks (keskkondadeks):

| Keskkond | Selgitus |
| --- | --- |
| Andmehõive ja töötluse keskkond | Autenditud kasutajaga keskkond andmete sisestamiseks Suurandmete süsteemi. Andmehõive keskkonna ülesanne on tagada andmete jõudmine Suurandmete süsteemi ning nende töötlemine teenuste jaoks vajalikule kujule.  Andmetöötlus võib toimuda, kas taustaprotsessidena või kasutaja poolt. Taustaprotsesside kasutamiseks planeeritakse eraldi server vastava tarkvaraga. |
| Analüütikakeskkond | Autenditud kasutajaga keskkond andmete analüüsimiseks ja kasutamiseks uuringute eesmärgil tuvastamaks andmete reaalset sisu ning kvaliteeti eesmärgiga valmistada ette uusi andmeteenuseid. Teenusteks võivad olla nii infoteenused riiklike registrite andmete osas, kui tuletatud andmeid väljastavad teenused, näiteks NPK ja huumusbilansi kalkulaatorid, väetussoovitused ja väetamiskaardid.  Analüütika keskkonna väljundiks on alusandmete ja andmete kvaliteedi visualisatsioonid ning failid (csv, xlsx), mis vormivad andmed tarbija jaoks sobivale kujule. Visualisatsioonid kujutavadki endast elektroonilisi teenuseid tarbijale. Ruumiandmete visualisatsioonid ja kaardifailid võivad olla põllutöömasinatele sisendiks tööde tegemisel. |
| Teenuste veebikeskkond | Keskkond jaguneb kaheks:   1. autenditud kasutajaga keskkond; 2. autentimata kasutajaga avaandmete keskkond.   Autenditud kasutajaga toimib isiku- ja põllupõhiste päringute tegemise keskkond, kus tootja saab infot enda ja oma ettevõtte objektide kohta, kasutades erinevaid Suurandmete süsteemi teenuseid. Kasutajale annab õigused Suurandmete süsteemi haldaja juhul, kui on olemas õiguslikud alused andmete kasutamiseks konkreetse isiku poolt.  Autentimata kasutajaga keskkonnas saab kasutaja vaadelda valdkonna avaandmeid ning kasutada avaandmete peale loodud teenuseid. |
| Suurandmete süsteemi haldamise keskkond | Autenditud kasutajaga keskkond, kus tegutsevad Suurandmete süsteemi haldajad. Peamised tegevused on:   1. Metaandmete ehk andmekirjelduste haldamine. Andmete kirjeldused sisaldavad mõõdikute ja dimensioonide kirjeldust ning andmebaaside ja teenuste andmemudeleid. 2. Klassifikaatorite ja koodilistide haldamine. 3. Süsteemi andmebaasi ja rakenduste tehniline haldamine. 4. Andmesisestusvormide häälestamine, testimine ja monitoorimine. 5. Suurandmete süsteemi teenuste häälestamine ja monitoorimine. 6. Kasutajate ja õiguste gruppide haldamine. 7. Reaalandmete andmehõive välistest allikatest, failidest ja käsitsi sisestades. |

**Tabel 2. Suurandmete süsteemi alamsüsteemid**

Tabelis on kirjeldatud süsteemi peamised alamsüsteemid.

## Andmehõive ja töötluse süsteem

Andmehõive on andmete liigutamine ka käsitsi või automaatsete vahenditega Suurandmete süsteemi. Suurandmete süsteem peab võimaldama andmeid lisaks kasutamisele ka hallata sh sisestada. Sellisteks andmeteks on näiteks info taimekahjustajate kohta, katsete tulemused (EMÜ) või agrometeoroloogia andmed, mis on praegu osaliselt paberil, Exceli tabelites või teksti kujul.

Andmete viimine suurandmete süsteemi andmebaasi võib toimuda kolmel moel:

1. elektroonsest allikast andmete hankimine ja viimine andmebaasi;
2. käsitsi andmete sisestamine;
3. andmefailide viimine andmebaasi.

Nimetatud võimalused peavad olema Suurandmete süsteemis toetatud. Allpool on kirjeldatud täpsemad nõuded nende tegevuste lahendamise osas.



**Joonis 1. Andmehõive ja töötluskeskkonna komponendid**

Kasutaja, kes ei ole süsteemi haldaja ja kes saab andmeid süsteemi edastada, näeb süsteemist reaalandmete sisestamise moodulit ja selle kaudu ka metaandmeid. Lisaks on kasutajal võimalus reaalandmete osas teha täiendavaid töötlusi ja arvutusi täiendamaks andmestikku lisanäitajatega.

Kõik tegevused andmehõive ja töötluskeskkonnas logitakse ning kasutajale nähtavad rakendused on monitooringuga kaetud.

Andmehõive keskkonna peamine omadus on, et see on haldusmoodulist kohandatav erinevate andmete hõive jaoks. Süsteemi abil on võimalik luua kasutajaliidesed nendele andmetele, millel täna andmekogu ja haldustarkvara puuduvad.

Järgnevas punktis on selgitatud mida tähendab andmete standardne vorm ja kuidas saavad sisult erinevad andmed olla samas struktuuris.

### Andmete standardne vorm

Üks suurtest probleemidest andmete käsitlemisel on nende erinev struktuur, defineerimata sisu ning kvaliteediprobleemid.

Suurandmete süsteemi salvestatavad andmed viiakse standardsele kujule. Kontseptsiooni aluseks on eeldus, et andmed tekivad mingi sündmuse käigus. Olgu selleks siis mullaanalüüs, ilmajaamas toimunud mõõtmine või saagi koguse mõõtmine kombainis. Tehakse erinevaid mõõtmisi kas käsitsi või automaatselt andurite abil. Ühe sündmusega võib tekkida 1-n näitajat, mis võivad olla nii numbrilised, kui ka koosneda tekstist. Näitaja sisu defineerib mõõdik. Sündmusega on alati seotud dimensioonid. Dimensioone võidakse täpsustada ka hiljem, kui sündmuse andmed on juba andmebaasis olemas. Erinevat liiki sündmuste andmekomplektid võivad olla erinevad. Sündmuse liigid tuleb defineerida klassifikaatorina. Selle elementide (sündmuse liikide) külge tuleb siduda konkreetsed mõõdikud. Peale reaalandme lisamist moodustub kokku andmestruktuur, mis sisaldab nii reaalandmeid, kui ka nende metaandmeid, mida kõike on võimalik koos ka süsteemist pärida ja kasutada visualisatsioonides nii kasutajaliidesega teenustes, kui ka veebiteenustes.

Andmete abstraktseks viidud struktuuri korral tuleb arvestada teatud riskidega. Sõltuvalt andmebaasimootorist võivad sellises süsteemis jääda päringud aeglaseks. Probleemi saab siiski lahendada andmete transformeerimise teel või kasutades paralleeltöötlust. Juhuks, kui on tarvis teha suuremahulisemaid analüüse, on võimalik andmestruktuur viia teisele kujule just analüüsitavate andmete osas – moodustada analüüsi andmebaas, mida on analüüsis mugavam ja efektiivsem kasutada. Nii moodustuvad andmetest tooted, mida on võimalik kasutada nii teenustes, kui ka analüüsikeskkonnas.

Transformeerimine tähendab alati ka ajalist viivitust andmete uuenemisel. Seda tuleb arvestada teenustes ja analüüsides.

Andmete töötlemise osas võib eristada järgmisi tegevusi:

1. **Andmete integreerimine ja transformeerimine** – mitme allika andmete kokku panemine, andmete viimine kasutamiseks sobivale kujule.
2. **Andmete klassifitseerimine ja kodeerimine** – tektsilistele andmete koodi omistamine, mis klassifitseerib andmekirje (sündmuse).
3. **Andmete valideerimine** – kvaliteedikontrolli läbiviimine ja selle tulemuste visualiseerimine.
4. **Andmete redigeerimine ja imputeerimine**  - juhul, kui andmetel esineb selliseid puuduvaid väärtusi, mis takistab nende kasutamist, siis luuaks nn imputeeritud andmed kas teistest andmetest tuletamise teel või mingi muu aktsepteeritud algoritmi alusel. Näiteks, kui on tarvis kokku saada mingi konkreetse mulla näitaja aegrida, et seda arvutustes kasutada, siis tuleb võtta viimati tehtud mulla analüüsi andmed ning arvutada järgmise perioodi näitaja välja arvestades põllul kasvatatud kultuuri, kasutatud väetise kogust ja sisaldust ning teisi näitajaid, mis veel on olulised puuduva väärtuse tuvastamiseks. Juhul, kui muid andmeid ei ole tuleb võtta näitaja eelmisest perioodis ning tekitada virtuaalne imputeeritud mullaanalüüsi sündmus. Selliselt saadud sündmuse info peab olema eristatud vastava tunnuse abil teistest näitajatest.
5. **Uute näitajate tuletamine** – olemasoleva sündmuste põhjal täiendavate näitajate arvutamine ja kirjeldamine.
6. **Andmete agregeerimine** – andmete summerimine üldistatud tasemele.
7. **Kaalude arvutamine** – näitajatele või summadele kaalude arvutamine, kui tegemist on valimi andmetega.
8. **Väljundi formeerimine** – töötluse tulemuse vormistamine andmebaasiks või andmefailiks.

Nimetatud tegevusi peab Suurandmete süsteem võimaldada sooritada. Töötluse algoritme peab saama häälestada (programmeerida) vastavalt vajadusele.



**Joonis 2. Standardne andmete struktuur erinevate andmestike jaoks**

Joonisel on kujutatud andmete haldamiseks peab Suurandmete süsteem vastama järgmistele nõuetele:

1. Andmete struktuurid defineeritakse sündmuse liigi klassifikaatori abil, kus igale liigile vastab kindel hulk mõõdikuid ning dimensioone, mis iseloomustavad sündmuste toimumise tingimusi.
2. Sündmuste liikide definitsioonide abil peab süsteem olema võimeline genereerima kasutajaliidese andmete sisestamiseks.
3. Süsteem peab võimaldama sündmuse ja näitajate andmeid importida XLSX ja CSV failide abil. Faili importimisel kontrollib süsteem andmete vastavust sündmuse liigi ja selle andmekomplekti definitsioonile.
4. Süsteem peab võimaldama salvestada andmebaasi kõiki ülal toodud joonisel näidatud andmeid. Tuleb arvestada, et osade andmete sisestamisel peab kasutama ka andmete sisestamise seadme GPS-i asukoha hinnangut ning asukoha koordinaate peab saama määrata seadmes oleva kaardi abil.
5. Keerukamate seostega andmete sisestamiseks luuakse tulevikus süsteemi eraldi andmesisestuse vormid, kuid andmed tuleb salvestada standardsesse andmestruktuuri.
6. Suurandmete süsteem peab olema võimeline andmeid materjaliseerima nii, et andmed denormaliseeritakse esimesel normaalkujul olevateks tabeliteks, kus ühe andmekomplekti andmed (sündmuse liigi kõikide mõõdikute ja dimensioonide näitajad) on ühes tabelis, iga lahter sisaldab ühte väärtust ning tulem väljastatakse ühe tabelina xlsx või csv formaadis kasutajale. Andmete väljastamisel peab kasutaja saama andmeid filtreerida sündmuse liigi, ajaperioodi, mõõdikute ja dimensioonide järgi nii, et sündmuse liigi koha väljastatakse kasutaja poolt soovitud veergude arv andmeid.
7. Süsteem peab võimaldama võtta andmeid süsteemist välja ühe mõõdiku põhiselt nii, et tulem on kõik reaalandmed üle kõikide andmeallikate.
8. Suurandmete süsteem peab võimaldama teostada andmete kvaliteedikontrolli.
9. Andmete osas peab saama viia läbi töötlusi mille käigus moodustatakse väljavõtted andmetest. Algandmed jäävad andmebaasi sellisena nagu need saabusid.

#### Näitajad

Tabelis NAITAJA on sooritatud mõõtmise tulemus ehk saadud näitaja, mis võib olla nii temperatuuri mõõtmine, saagi mahu või kaalu mõõtmine, väetise koguse mõõtmine või ka mingile nõudele vastavuse kontrollimine.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Primaar-  võti | Atribuudi nimi | Andme-  tüüp | Kohustus-  likkus | Kirjeldus |
| True | ID | bigint | True | Mõõtmise unikaalne identifikaator. |
| False | MOODIK\_KOOD | varchar(100) | True | Viide mõõdikule. Mõõdiku juures on kirjeldatud, milline on täpsem näitaja sisu. Mõõdiku kaudu on kättesaadav kogu metainfo alates mõõdiku sisulisest selgitusest, kuni klassifikaatoriteni välja. |
| False | VAARTUS\_TEKST | varchar(max) | True | Väärtus võib olla kas konkreetne mõõmisel saadud väärtus või siis viide klassifikaatori elemendile, kui mõõtmise tulemus on väljendatav skaalal, mis on kirjeldatud klassifikaatorina (koodilistina). Näiteks kvaliteedi mõõtmisel võivad olla väärtusteks "Madal", "Keskmine" või "Kõrge". Teksti kujul võib siin olla ka mingi pikem struktureerimata väärtus, näiteks mõõtmisel saadud veateade. |
| False | SYNDMUS\_ID | bigint | True | Viide sündmusele. |
| False | TULETATUD | bit | True | Näitab, kas näitaja andmed on tuletatud teiste andmete põhjal. Tuletatakse hädavajadusel, kui puuduv info takistab arvutusi, päringute tegemist või teenuse osutamist. |

**Tabel 3. Näitajate andmestruktuur**

#### Sündmused

Tabelid SYNDMUS registreeritakse sündmus, mille käigus toimus mingite näitajate tuvastamine. Juhul, kui on tarvis lisada tekstilisi selgitusi sündmuse või näitaja juurde, siis tuleb kasutada näitajat ja mõõdikut, mis väljendavad selgitust.

| Primaar-  Võti | Atribuudi nimi | Andme-  tüüp | Kohustus-  likkus | Kirjeldus |
| --- | --- | --- | --- | --- |
| True | ID | bigint | True | Sündmuse unikaalne identifikaator. |
| False | SYNDMUS\_LIIK | varchar(100) | True | Viide sündmuse liigile. Sündmuse liik tuleb defineerida klassifikaatorina. Sündmuse võimalike näitajate komplekt tuleb defineerida samuti klassifikaatorina, kus sündmuse liigi ja andmeelemendi vahele tuleb tekitada vastav sündmuse andmekomplekti seos. |
| False | AEG | datetime | True | Sündmuse toimumise aeg. |
| False | X-KOORDINAAT | varchar(50) | False | Mõõtmise teostamise asukoha X-koordinaat. |
| False | Y-KOORDINAAT | varchar(50) | False | Mõõtmise teostamise asukoha Y-koordinaat. |
| False | Z-KOORDINAAT | varchar(50) | False | Mõõtmise asukoha kõrgu merepinnast. |
| False | KOORDINAATSYSTEEM | varchar(50) | False | Koordinaatsüsteemi tähis vastavalt ISO, milles koordinaadid on väljendatud EPSG:nnnn kujul. Vt. https://spatialreference.org/ref/epsg/ . |
| False | AADRESS\_ID | bigint | False | Viide mõõtmiskoha aadressile. |
| False | SYNDMUS\_FAIL\_ID | bigint | False | Faili ID, millest on pärit sündmuse andmed. |
| False | TULETATUD | bit | True | Näitab, kas sündmuse andmed on tuletatud teiste andmete põhjal. Tuletatakse hädavajadusel, kui puuduv sündmus takistab arvutusi, päringute tegemist või teenuse osutamist. |
| False | KIRJE\_LISATUD | timestamp | True | Millal kirje lisati. |
| False | KIRJE\_MUUDETUD | timestamp | False | Millal kirje muudeti. |

**Tabel 4. Sündmuste andmestruktuur**

#### Sündmuse ja dimensiooni seos

Tabelisse SYNDMUS\_DIMENSIOON salvestatakse sündmuse ja dimensiooni seosed. Näiteks milliseid aineid seoses sündmusega kasutati või millist seadet.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Primaar-  võti | Atribuudi nimi | Andme-  tüüp | Kohustus-  likkus | Kirjeldus |
| False | SYNDMUS\_ID | bigint | True | Viide mõõtmisE sündmusele. |
| False | DIMENSIOON\_KOOD | varchar(100) | True | Viide dimensioonile. |
| False | VAARTUS\_TEKST | varchar(max) | False | Milline oli dimensiooni väärtus sündmuse hetkel. Näiteks milline oli põllu identifikaator. |
| False | LISAMINE\_AEG | timestamp | True | Sündmusele dimensiooni lisamise aeg. |

**Tabel 5. Sündmuse ja dimensiooni seose andmestruktuur**

#### Sündmuste andmefailid

Tabelisse SYNDMUS\_FAIL salvestatakse mõõtmise tulemusel tekkinud fail(id).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Primaar-  võti | Atribuudi nimi | Andme-  tüüp | Kohustus-  likkus | Kirjeldus |
| True | ID | bigint | True | Faili unikaalne identifikaator. |
| False | NIMI | varchar(2000) | True | Faili nimi algallikas. |
| False | SISU\_BASE64 | varbinary(max) | True | Faili sisu base64 kodeeringus. |
| False | MIMETYPE | varchar(50) | True | Faili tüüp. |
| False | KOMMENTAAR | varchar(4000) | False | Kommentaar faili sisu kohta. |

**Tabel 6. Sündmustega seotud andmefailide andmed**

## Paralleelprotsesside jooksutamine taustal

Suuremad andmete analüüsid ja töötlused võivad võtta aega väga kaua. Nende jaoks tuleb ette näha eraldi server ja klaster, kus on võimalik nii andmeid, kui protsesse jagada tükkideks ning protsesse käivitada paralleelselt. Suuremateks andmeteks, mille puhul juba tekib paralleelse töötluse vajadus on näiteks ilmastiku andmed, Maa-ameti poolt lidariga kaardistatud maapinna kõrgusmudel, samuti masinatest saabuvad või masinasse minevad andmed, kui neid kasutada algsel kujul. Samuti satelliidipiltide põhjal loodud indeksid, kui nende eraldusvõime on 1-3m küljepikkusega ruutude tasemele viidud.

Paralleeltöötlus nõuab töötluse programmeerimist vastavas keskkonnas. Erinevatest võimalustest on juttu arhitektuurivariantide peatükis.

## Analüütikakeskkond

Analüütikakeskkond on mõeldud:

1. teadlastele andmete analüüsimiseks;
2. statistikutele andmete analüüsimiseks;
3. teenuste ettevalmistamiseks ja andmetele hinnangute andmiseks kvaliteedi ning sisu osas;
4. süsteemide arendajatele andmete tutvustamiseks ja lahenduste katsetamiseks.

Analüütika keskkonna jaoks ei ole mõtet luua tarkvara nullist vaid mõistlik on kasutada olemasolevaid valmis analüütikatarkvarasid. Analüütikakeskkonna komponente iseloomustab järgnev joonis:



**Joonis 3. Analüüsikeskkonna komponendid**

Joonisel on kujutatud üks võimalikest analüüsikeskkonna konfiguratsioonidest. Analüüsikeskkond peab olema võimalik liidestada andmeallikatega läbi adapterite. Levinumate analüüsitarkvarade koosseisu kuuluvad adapterid, mis toetavad laiemalt levinud andmebaase (Oracle, Postgres jt). Samuti on toetatud veebiteenused. Veebiteenuste osas on Suurandmete süsteemis võetud kasutusele Open Data standard (vt. [[5](https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=odata)]). Ruumiandmete osas kasutatakse WMS, WFS ja WCS teenuseid. Eelistatud ühendumise mooduseks ongi veebiteenuste kasutamine.

Lisaks välistele ühendustele kasutatakse Suurandmete süsteemi siseseid andmeid, milledeks on metaandmed ja reaalandmed nähtuste kohta, mille jaoks eraldi andmekogu seni pole loodud. Andmekogu rolli hakkab nende andmete osas mängima Suurandmete süsteem.

Analüütikul on võimalik analüüsikeskkonda kasutada, kas veebilehitsejas või sisenedes analüüsi keskkonda ja kasutades seal *desktop* rakendust. Analüüsitarkvarade eripäraks on suuremate analüüsi tegemisel mastaapne mälukasutus sest arvutusi tehakse mällu laetud andmete peal. See tähendab, et arvuti mälukasutus tõuseb kergesti üle 8GB, mis on tavakonfiguratsiooniga arvutis. Analüüsi teostavas masinas ei tohiks olla mälu alla 32GB. Efektiivsema ressursikasutuse tagamiseks ongi mõtet luua analüütikule virtuaalne *desktop* masin, kus on olemas piisav mäluressurss.

Erinevate analüüside ja visualiseerimise jaoks tuleb tihti kasutada mitut rakendust. Näiteks ruumiandmete analüüsimiseks ei pruugi sobida levinud raporteerimise (BI) jaoks mõeldud tarkvarad ja vastupidi – ruumiandmete analüüsi tarkvarad (GIS) võivad hätta jätta aruandluse koostamisel.

Suurandmete süsteemi analüütikatarkvara osas kasutatakse valmis tarkvarapakette, millel on suur hulk valmis arendatud adaptereid ja visualiseerimise tööriistu. Puuduva funktsionaalsuse arendamine peab olema võimalik nii adapterite, analüüsipakettide kui ka visualiseerimiskomponentide osas. Arenduste lähtekood peab olema avatud sh adapterite oma. Keerukamate analüüside jaoks peab olema võimalik kasutada laialt levinud R-i ja Pythoni skripte. Kasutusele võetav analüüsitarkvara ei tohi seada takistusi täiendavate tarkvarade kasutuselevõtuks. Suurandmete konfiguratsioon sisaldab teatud tarkvarasid aga analüütikud võivad kasutada ka muid tarkvarasid, millel on olemas analoogsed adapterid.

## Teenuste veebikeskkond

Suurandmete süsteemi teenuste keskkonna kasutajateks on põllumajandustootjad ning teised õigustatud huviga isikud, kelle jaoks on mõeldud autentimisega keskkond, mille kaudu on võimalik juurde pääseda isikuga seotud andmetele.

Lisaks isikustatud andmetele on teenuseid, mis ei nõua autentimist. Näiteks erinevate avalike registrite info jagamise teenused.



**Joonis 4. Suurandmete süsteemi teenuste keskkond**

Joonisel on kujutatud teenuste keskkonna põhimõtteline skeem, kus pole välja toodud kõiki andmeallikaid ega teenuseid. Suurandmete süsteemi teenustest on täpsemalt juttu teenuste peatükis.

## Suurandmete süsteemi haldamise keskkond

Suurandmete süsteemi haldaja kasutab eraldi keskkonda Suurandmete töö juhtimiseks.



**Joonis 5. Suurandmete süsteemi haldamise keskkond**

Haldamise lahendus jaguneb järgnevateks alamkeskkondadeks:

1. Paralleeltöötluse moodul – moodul, kus on võimalik programmeerida ja käivitada andmete paralleeltöötlusi ja analüüse. Tegemist on andmete haldusest eraldi seisva lahendusega.
2. Andmehalduse moodul – mõeldud süsteemi metaandmete ja Suurandmete süsteemi koosseisu kuuluvate reaalandmete haldamiseks.
3. Monitooringu moodul – siin on võimalik seadistada kõikide Suurandmete süsteemi teenuste monitooringut ning jälgida monitooringust laekuvaid andmeid. Lisaks on siin ka logide salvestamine. Logimise osas on tarvis eraldi teenust, mida teised alamsüsteemid kasutavad. Logid salvestatakse tekstifailidesse, et need oleksid rakendustest sõltumatud.

Halduskeskkonnale esitatavad nõuded on täpsemalt kirjeldatud nõuete peatükis.

## Suurandmete süsteemi nõuete loetelu

Järgnevas tabelis on toodud analüüsikeskkonnale esitatavad funktsionaalsed ja mittefunktsionaalsed nõuded.

| ID | Funktsionaalsuse grupp | Funktsionaalsus (nõue) | Nõude kirjeldus |
| --- | --- | --- | --- |
| REQ-1 | 1. Metaandmete haldus | Klassifikaatorite haldus | Klassifikaatorit peab olema võimalik süsteemi lisada, sisestades klassifikaatori koodi nimetuse ja muud atribuudid, mis on kirjeldatud andmemudelis. Samuti peab olema võimalik kõiki neid muuta va kood. Koodi muutmine tähendab uue klassifikaatori lisamist. |
| REQ-2 | 1. Metaandmete haldus | Klassifikaatori versioonide haldus | Klassifikaatorist peab saama moodustada uut versiooni. Versiooni moodustamine tähendab sama koodiga uue klassifikaatori moodustamist, mille käigus lisatakse sisuliselt uus klassifikaator ja eelmise versiooni elementidest tehakse koopia. Klassifikaatori versiooni ei saa teha aegpidevast klassifikaatorist. |
| REQ-3 | 1. Metaandmete haldus | Klassifikaatorite võrdlemine | Süsteem peab võimaldama võrrelda kahte klassifikaatorit. Võrdluse tulemus on klassifikaatori väljade väärtuste võrdlus ja klassifikaatori elementide võrdlus, milles tuuakse välja sarnasused ja erinevused. Võrdlemine peab olema nö tark ja suutma siduda väikese kirjapildi erinevusega klassifikaatori elemente. Võrdlus peab võrdlema elemente nii nimede, kui koodide järgi ja juhul, kui koodid ei kattu, kuid nimed kattuvad, näitama antud seose sarnasusena ära. |
| REQ-4 | 1. Metaandmete haldus | Klassifikaatorite elementide haldus | Süsteem peab võimaldama klassifikaatorite elementide lisamist, muutmist ja kustutatuks märkimist. Elemendi atribuutide hulk on määratud andmemudelis. |
| REQ-5 | 1. Metaandmete haldus | Klassifikaatori elementide versioonide haldus | Süsteem peab võimaldama hallata klassifikaatori elemendi versioone nii, et juhul, kui elemendi andmeid muudetakse salvestatakse eelmine seis eraldi kirjena ja märgitakse kehtetuks. |
| REQ-6 | 1. Metaandmete haldus | Elementide vaheliste seoste haldus | Süsteem peab võimaldama moodustada klassifikaatorite elementidest nii hierarhiad, kui ka võrgustikke. Võrgustik tähendab seda, et ühel elemendil võib olla 1..n vanemat. Hierarhia moodustub juhul, kui igal elemendil on 0..1 vanem ja 1..n last. Siduda peab saama erinevate klassifikaatorite elemente nii, et oleks võimalik koostada seostabeleid, mille alusel on võimalik andmestikke ühest klassifikaatori alt teise ümber kodeerida. |
| REQ-7 | 1. Metaandmete haldus | Elemendi seosetüüpide haldus | Süsteemis peab olema võimalik hallata elementide vaheliste seoste tüüpe. Alati, kui klassifikaatori elemendi seos luuakse määratakse ka seose suund ja tüüp (vanem, laps, osa jne). |
| REQ-8 | 1. Metaandmete haldus | Elementide järjestamine | Süsteemis peab olema võimalik määrata klassifikaatori elementidele järjestusi ühe klassifikaatori piires. Ühe klassifikaatori piires võib elementidel olla mitu järjestust. |
| REQ-9 | 1. Metaandmete haldus | Elementide dünaamiline struktuur, atribuutide lisamine | Klassifikaatori elemendi atribuutide (andmeväljade) arv ei tohi olla piiratud. Süsteemis peab olema võimalik lisaks kohustuslikele elemendi atribuutidele, lisada klassifikaatori elemendile suvalise arvu andmevälju. Andmeväljade koosseis kehtid ühe klassifikaatori piires. |
| REQ-10 | 1. Metaandmete haldus | Klassifikaatori kopeerimine | Süsteem peab võimaldama kopeerida klassifikaatori sisu nii, et moodustatakse uus klassifikaator kopeeritava klassifikaatori sisuga, kui uue klassifikaatori koodiga. |
| REQ-11 | 1. Metaandmete haldus | Klassifikaatori struktuuri kopeerimine | Klassifikaatori elementide andmekoosseisu kirjeldust peab saama aluseks võtta uue klassifikaatori moodustamisel. Siin tuleb eristada klassifikaatori kopeerimist ja klassifikaatori struktuuri kopeerimist. |
| REQ-12 | 1. Metaandmete haldus | Andmemudelite kirjeldamine, UML | Süsteemis peab olema võimalus hallata andmestruktuuri kirjeldust UML andmemudelina (ERD). |
| REQ-13 | 1. Metaandmete haldus | Andmemudelite import ja eksport | Süsteem peab võimaldama andmemudeleid (ja ka teisi UML mudeleid) süsteemi importida ja süsteemist eksportida. Andmevahetusformaat on XMI uusim versioon (http://www.omg.org/spec/XMI/2.5.1). |
| REQ-14 | 1. Metaandmete haldus | Andmemudeli pöördprojekteerimine andmebaasist, Oracle | Andmemudelite haldamise süsteem peab võimaldama pöördprojekteerida andmebaasi mudelit Oracle andmebaasi pealt. See tähendab andmemudeli uuendamist andmebaasi põhjal. |
| REQ-15 | 1. Metaandmete haldus | Andmemudeli pöördprojekteerimine andmebaasist, Potgres | Andmemudelite haldamise süsteem peab võimaldama pöördprojekteerida andmebaasi mudelit Postgres andmebaasi pealt. See tähendab andmemudeli uuendamist andmebaasi põhjal. |
| REQ-16 | 1. Metaandmete haldus | Andmemudelite kirjeldamine UML andmemudelina | Süsteem peab võimaldama luua ja hallata andmudeleid UML kujul. |
| REQ-17 | 1. Metaandmete haldus | Andmebaasi loomise skripti (DDL) genereerimine | Süsteem peab võimaldama andmebaasi loomise skripti genereerimist andmemudeli pealt selliselt, et skript on käivitatav andmebaasiserveris. |
| REQ-18 | 1. Metaandmete haldus | Andmeallikate kirjeldamine | Süsteem peab võimaldama hallata andmeallikate metaandmeid (vt andmemudel). |
| REQ-19 | 1. Metaandmete haldus | Andmeallika omaduste kirjeldamine | Süsteem peab võimaldama kirjeldada andmeallikate metaandmeid teenuste osas nii, et kirjeldusest oleks võimalik kindlaks teha, kust ning kuidas on võimalik teenust kasutada ning mis on teenuse sisu. |
| REQ-20 | 1. Metaandmete haldus | Mõõdikute kirjeldamine | Süsteem peab võimaldama mõõdikute kirjelduste haldamist. Mõõdiku on andmeelement, mille väärtus iseloomustab mingi konkreetse objekti või nähtuse mingit omadust, mis on mõõdikuga iseloomustatav. |
| REQ-21 | 1. Metaandmete haldus | Mõõdikute vaheliste seosvalemite kirjeldamine | Süsteem peab võimaldama kirjeldada mõõdikute vahelisi seoseid sh mõõdikute seosvalemeid. |
| REQ-22 | 1. Metaandmete haldus | Dimensioonide kirjeldamine | Süsteem peab võimaldama kirjeldada dimensioone. Dimensoonide all peame silmas andmeobjekti iseloomustavaid mittemõõdetavaid aspekte nagu asukoht, liik jne. |
| REQ-23 | 1. Metaandmete haldus | Mõõdikute ja dimensioonide vaheliste seoste kirjeldamine | Süsteem peab võimaldama hallata dimensioonide ja mõõdikute vahelisi seoseid. Antud seos näitab, millistel tingimustel (dimensioon) konkreetne näitaja (mõõdiku väärtus) ilmnes. |
| REQ-24 | 2. Andmete integratsioon | Andmestruktuuri kirjeldamine | Süsteem peab sisaldama andmesisestusmoodulit, mille struktuur on dünaamiline (haldaja poolt muudetav) nii, et oleks võimalik tekitada andmestruktuure erinevate andmete sisestamiseks (näiteks taimekahjustajate tuvastamise andmed, mille üheks osaks on pilt, mis põllul tehti). |
| REQ-25 | 2. Andmete integratsioon | Sisestusvormide kirjeldamine | Süsteem peab sisaldama andmete sisestamise vormide kirjeldamise võimalust nii, et süsteemi haldajal oleks võimalik kasutaja jaoks luua erinevaid vorme andmete edastamiseks serverisse. |
| REQ-26 | 2. Andmete integratsioon | Kontrollide kirjeldamine | Andmete sisestamise vormidel peab olema võimalik kirjeldada kontrolle, mis rakenduvad andmete sisestamisel. Kontrollida peab olema võimalik konkreetset väärtust, väärtuste vahemikku numbrite korral, väärtuse puudumist, tingimuslikku väärtuse puudumist sh kasutades kõiki aritmeetilisi operatsioone. |
| REQ-27 | 2. Andmete integratsioon | REST teenuse kasutamine | Süsteem peab võimaldama andmesisestusvormil kasutada REST teenuseid andmete pärimiseks, mida kasutaja saab sisestusvormil kasutada. |
| REQ-28 | 2. Andmete integratsioon | X-tee teenusest andmete kasutamine | Süsteem peab võimaldama siduda andmete sisestamise vormi X-tee teenusega nii, et kasutaja poolt sisestatud andmete peale kutsub süsteem X-tee teenust. Näiteks, kui kasutaja sisestab isikukoodi, siis süsteem pärib Rahvastikuregistrist isiku andmed ning kuvab need ekraanil. |
| REQ-29 | 2. Andmete integratsioon | WFS veebiteenustest andmete kasutamine | Süsteem peab võimaldama kasutada andmesisestusvormidel geoinfot WFS teenusest. Näiteks põllumassiivide registri põllumassiivide infot. |
| REQ-30 | 2. Andmete integratsioon | WCS veebiteenustest andmete kasutamine | Süsteem peab võimaldama kasutada andmesisestusvormidel geoinfot WCS teenusest. |
| REQ-31 | 2. Andmete integratsioon | WMS veebiteenustest andmete kasutamine | Süsteem peab võimaldama kasutada andmesisestusvormidel geoinfot WMS teenusest. |
| REQ-32 | 2. Andmete integratsioon | Andmebaasist andmete kasutamine (Oracle) | Süsteem peab võimaldada kasutada andmeid Oracle andmebaasist. |
| REQ-33 | 2. Andmete integratsioon | Andmebaasist andmete kasutamine (Postgres) | Süsteem peab võimaldama kasutada andmeid Postgres andmebaasist. |
| REQ-34 | 2. Andmete integratsioon | CSV andmehõive (import) | Süsteem peab võimaldama kasutada andmeid CSV failist (andmete import). |
| REQ-35 | 2. Andmete integratsioon | XLSX andmehõive (import) | Süsteem peab võimaldama kasutada andmeid XLSX failist vähemalt ühe töölehe ulatuses ilma mahupiiranguta. |
| REQ-36 | 2. Andmete integratsioon | Andmebaasi andmete salvestamine (Postgres) | Süsteem peab võimaldama salvestada andmeid Postgres andmebaasi. |
| REQ-37 | 2. Andmete integratsioon | GPS andmete pärimine ja kasutamine | Süsteemiga defineeritavad andmesisestusvormid peavad võimaldada kasutajal kasutada automaatselt kasutaja seadmes olevat GPS andurit nii, et anduri poolt antav asukoht salvestud kasutaja poolt sisestatud andmete juurde. |
| REQ-38 | 2. Andmete integratsioon | Seadmest fotode kasutamine | Süsteem peab võimaldama sisestusvormi kaudu siduda kasutaja poolt sisestatud andmetega kasutaja poolt tehtud fotosid, mis on kasutaja seadmest kättesaadavad. |
| REQ-39 | 2. Andmete integratsioon | Seadmest failide kasutamine | Süsteem peab võimaldama sisestusvormi kaudu siduda kasutaja poolt sisestatud andmetega kasutaja seadmele kättesaadavaid faile. |
| REQ-40 | 2. Andmete integratsioon | Otsepäringute kasutamine | Süsteem peab võimaldama kasutada andmeallika andmeid reaalajas ilma neid kopeerimata või puhverdamata. Kasutajal peab olema võimalus seadistada andmekasutus kas puhverdamisega või reaalajas. |
| REQ-41 | 2. Andmete integratsioon | REST teenusest andmete pärimine | Süsteem peab võimaldama kasutada andmeid REST teenustest. |
| REQ-42 | 2. Andmete integratsioon | X-tee teenusest andmete pärimine | Süsteem peab võimaldama kasutada andmeid X-tee teenutest. |
| REQ-43 | 2. Andmete integratsioon | WFS veebiteenustest andmete pärimine | Süsteem peab võimaldama andmeid kasutada WFS teenustest. |
| REQ-44 | 2. Andmete integratsioon | WCS veebiteenustest andmete pärimine | Süsteem peab võimaldama andmeid kasutada WCS teenustest. |
| REQ-45 | 2. Andmete integratsioon | WMS veebiteenustest andmete pärimine | Süsteem peab võimaldama andmeid kasutada WMS teenustest. |
| REQ-46 | 2. Andmete integratsioon | Andmebaasist andmete pärimine, ODBC | Süsteem peab võimaldama kasutada andmeid ükskõik millisest andmebaasist, millel on olemas ODBC või JDBC ühenduse võimalus. |
| REQ-47 | 2. Andmete integratsioon | CSV andmehõive | Süsteem peab võimaldama kasutada andmeid CSV failist. |
| REQ-48 | 2. Andmete integratsioon | XLSX andmehõive | Süsteem peab võimaldama kasutada andmeid XLSX failist. |
| REQ-49 | 2. Andmete integratsioon | Andmebaasist andmete pärimine, Postgres | Süsteem peab võimaldama kasutada andmeid otse Postgres andmebaasist. |
| REQ-50 | 2. Andmete integratsioon | Andmebaasist andmete pärimine, Oracle | Süsteem peab võimaldama kasutada andmeid otse Oracle andmebaasist. |
| REQ-51 | 2. Andmete integratsioon | Andmebaasist andmete, Hadoop HDFS | Süsteem peab võimaldama kasutada andmeid otse HDFS-st. |
| REQ-52 | 2. Andmete integratsioon | Failide hõivamine | Süsteem peab võimaldama transportida faile ja neid paigutada süsteemi serverisse. |
| REQ-53 | 2. Andmete integratsioon | Sõnumite lugemine CAN BUS-lt (ISO BUS) | Süsteem peab võimaldama lugeda andmeid põllumajandusmasinates kasutatavalt CAN siinilt (ISO 11898). |
| REQ-54 | 2. Andmete integratsioon | Sõnumitele andmeanalüüsi rakendamine | Süsteem peab võimaldama töödelda CAN siinilt pärievat infot loetuna otse arvuti COM pordilt ilma vahetarkvaradeta va operatsioonisüsteem. |
| REQ-55 | 2. Andmete integratsioon | Sõnumite filtreerimine | Süsteem peab võimaldama CAN siini sõnumeid filtreerida nii, et on võimalik kätte saada ainult kasutaja poolt soovitud sõnumid. |
| REQ-56 | 2. Andmete integratsioon | Sõnumiahelate moodustamine | Süsteem peab võimaldama CAN siini sõnumitest moodustada ahelaid, sidudes omavahel kokku sõnumite jada, millede signatuurid on ette seadistatud. |
| REQ-57 | 3. Andmetöötlus | Töötluse programmeerimine | Süsteem peab võimaldama moodustada andmete hõive ja töötlemise pakette ning moodustada nendest töövooge. |
| REQ-58 | 3. Andmetöötlus | Töötluse testimine | Süsteem peab võimaldama koostatud andmetöötluse pakette ning töövooge testida. |
| REQ-59 | 3. Andmetöötlus | Töötluse ajastamine | Süsteem peab võimaldama kasutaja poolt koostatud pakette ja töövooge käivitada ajastatult. |
| REQ-60 | 3. Andmetöötlus | Töötluse perioodiline ajastamine | Süsteem peab võimaldama pakettide ja töövoogude käivitamist ajastada perioodiliseks. |
| REQ-61 | 3. Andmetöötlus | Töötluste versioonihaldus | Süsteem peab säilitama pakettide ja töövoogude muutmise ajaloo (versioneerimine). |
| REQ-62 | 3. Andmetöötlus | Töötluse tühistamine | Süsteem peab võimaldama paketi poolt tehtud töötlust tühistada, kui sellele pole järgnenud ühtki uut töötlust. |
| REQ-63 | 4. Andmeanalüüs | Automaatprofileerimine | Süsteem peab võimaldama andmete automaatprofileerimist, mis tooks välja veeru kohta väärtuste sagedused, tühjade väärtuste arvu, maksimum- ja miinimumväärtused. |
| REQ-64 | 4. Andmeanalüüs | Adhoc analüüsid (päringud, päringute genereerimine) | Andmeanalüüsitarkvara peab võimaldama teha andmete pealt vabalt valitud sisuga päringuid nii, et analüütik ei pea ise päringu SQL-i kootama vaid analüüsitarkvara genereerib selle. Siiski peab olema võimalus genereeritud päringu käsitsi muutmiseks. |
| REQ-65 | 4. Andmeanalüüs | Ruumiandmete analüüs ja visualiseerimine | Süsteem peab võimaldama teostada operatsioone ruumiandmetega sh paigutada andmeid kaardile aadressi ja geograafiliste koordinaatide järgi, leida ruumikuju pindala, leida ruumikuju keskpunkti koordinaati, leida ja kuvada etteantud suurusega puhvertsoone, leida ruumikujude kattuvuse suurust, leida kas üks kuju asub teise sees ja leida ruumikuju naabriteks olevaid ruumikujusid. |
| REQ-66 | 4. Andmeanalüüs | Profileerimistulemuste salvestamine | Süsteem peab võimaldama andmete profileerimise tulemuste salvestamist nii, et profiili kuvamisel ei tule alati kõike uuesti arvutada. |
| REQ-67 | 4. Andmeanalüüs | Lihtkontrollide kirjeldamine | Süsteem peab võimaldama vaadeldavate andmete suhtes defineerida andmekontrolle selliselt, et kontrolle ei tuleks kirjutada SQL-is vaid need oleksid sarnased kontoritarkvarades (näit. Excel) kasutatavatele keelele, mis on loetav ja arusaadav keskmisele arvutikasutajale. Minimaalselt peavad toetatud olema kõik aritmeetilised tehted, stringioperatsioonid tekstidest info kätte saamiseks ning astendamine. |
| REQ-68 | 4. Andmeanalüüs | Keerukate üle mitme andmekogu ja tabeli kontrollide kirjeldamine | Süsteem peab toetama andmete ristkasutamist ja kontrollide rakendamist üheaegselt mitme tabeli ja mitme andmeallika andmetele. |
| REQ-69 | 4. Andmeanalüüs | Kontrollitulemuste salvestamine | Süsteem peab võimaldama andmete kontrollimise tulemusi salvestada nii, et tulemuste vaatamisel neid igakord uuesti ei arvutata. |
| REQ-70 | 4. Andmeanalüüs | Kvaliteediaruande genereerimine | Süsteem peab võimaldada visualiseerida andmete kontrollimise tulemusi nii, et kuvatakse statistiline kokkuvõte ja lisaks kõik kirjed, mis kontrolle ei läbinud. Peab olema võimalik kirjeid kontrolli kaupa sirvida. |
| REQ-71 | 4. Andmeanalüüs | Kvaliteediaruande jagamine veebis | Süsteem peab võimaldama kvaliteediaruannet jagada veebilehena. |
| REQ-72 | 4. Andmeanalüüs | Andmemudeli koostamine üle mitme allika | Süsteem peab võimaldama koostada analüüsi aluseks olevat andmudelit üle mitme andmeallika ja sh üle mitme eritüübilise andmeallika. |
| REQ-73 | 4. Andmeanalüüs | Transformatsioonipäringute kirjutamise võimalus | Süsteem peab võimaldama teostada andmete transformeerimist sobivale kujule nii, et andmetabeleid pannakse kokku või lahutatakse, veerge liidetakse üheks ja lahutatakse mitmeks. |
| REQ-74 | 4. Andmeanalüüs | Sageduste arvutamine | Süsteem peab võimaldama arvutada väärtuste kasutamise sagedusi. |
| REQ-75 | 4. Andmeanalüüs | Klasterdamine erinevate tunnuste alusel | Süsteem peab sisaldama klasterdamise funktsionaalsust. |
| REQ-76 | 4. Andmeanalüüs | Filtreerimine | Süsteem peab võimaldama andmete interaktiivset filtreerimist nii analüüside ja visualisatsoone luues, kui ka visualisatsioone kasutades. |
| REQ-77 | 4. Andmeanalüüs | Ennustamine (forecast) | Süsteem peab sisaldama andmete põhjal prognoosimise funktsionaalsust. Süsteemi peab kasutaja saama juurde luua prognoosimise funktsionaalsust. |
| REQ-78 | 4. Andmeanalüüs | Analüüside programmeerimine | Süsteem peab võimaldama koostada andmete analüüsimise pakette kasutades R-i ja Pythonit. |
| REQ-79 | 4. Andmeanalüüs | R tugi (analüüside programmeerimine) | Süsteem peab andmeanalüüside jaoks võimaldama kasutada R-i. |
| REQ-80 | 4. Andmeanalüüs | Andmebaaside ja veebiteenuste adapterite kasutamise võimalus | Süsteem peab sisaldama adaptereid andmebaasidest ja veebiteenustest andmete kätte saamiseks. Olemas peavad olema vähemalt Oracle, Postgres, MS SQL Server, Access andmebaaside, Open Data veebiteenuste, XLSX, CSV, JSON ja XML failide ning ODBC adapterid. |
| REQ-81 | 4. Andmeanalüüs | In-memory analüüsi sooritamine | Süsteem peab võimaldama käivitada analüüse arvuti mälus nii, et kõik analüüsitavad andmed loetakse korraga mällu. |
| REQ-82 | 4. Andmeanalüüs | Analüüsi käivitamine rakendusserveril | Süsteem peab võimaldama käivitada analüüse rakendusserveril. Näiteks R-i serveril. |
| REQ-83 | 4. Andmeanalüüs | Analüüsi käivitamine andmebaasis | Süsteem peab võimaldama käivitada analüüse andmebaasis SQL päringuna. |
| REQ-84 | 4. Andmeanalüüs | Visualisatsioonide loomine | Süsteem peab võimaldama luua andmete põhjal interaktiivseid visualisatsioone. |
| REQ-85 | 4. Andmeanalüüs | Visualisatsiooni mallide loomine | Süsteem peab võimaldama koostada visualisatsioonide malle. |
| REQ-86 | 4. Andmeanalüüs | Relatsiooniliste andmete visualiseerimine | Süsteem peab võimaldama analüüsida ja visualiseerida relatsioonilisi andmeid nii, et automaatselt visualisatsioonid arvestavad automaatselt andmebaasi seoseid ja üht tabelit filtreerides filtreeritakse ka teist seotud tabelit. |
| REQ-87 | 4. Andmeanalüüs | Punktandmete visualiseerimine kaardil | Süsteem pea võimaldama visualiseerida andmeid kaardil geograafiliste koordinaatide abil. Toetatud peab olema vähemalt WGS84 koordinaatsüsteem. |
| REQ-88 | 4. Andmeanalüüs | Ekstrapoleerimine | Süsteem peab toetama andmete ekstrapoleerimist. Näiteks, kui põllu ühe või mõne punkti kohta on teada mingi väärtus, siis peab saame neid väärtusi kaardil laiendada kogu põllule nii, et näiteks visuaalis kuvatakse punktide lähedane ala kuni põllu piirini väärtusele iseloomuliku värviga. |
| REQ-89 | 4. Andmeanalüüs | Analüüsi tulemuste visualiseerimine kaardil | Süsteem peab võimaldama visualiseerida kaardil andmeid aadressi ja geograafiliste koordinaatide abil. Toetatud peab olema ruumikuju kasutamine ja kuvamine. |
| REQ-90 | 4. Andmeanalüüs | Erinevate kaardikihtide üheaegne kasutamine | Süsteem peab võimaldama kasutada mitut kaardikihti korraga. |
| REQ-91 | 4. Andmeanalüüs | Erinevate aluskaartide kasutamine | Süsteem peab võimaldama kasutada erinevaid aluskaarte. |
| REQ-92 | 4. Andmeanalüüs | Kaardikihtide pärimine veebiteenusest | Süsteem peab võimaldama kasutada kaardikihte WMS, WFS ja WCS teenustest. |
| REQ-93 | 4. Andmeanalüüs | Aruannete koostamine | Süsteem peab võimaldama andmeid väljastada aruande vormis. |
| REQ-94 | 4. Andmeanalüüs | Aruannete ligipääsuõiguste määramine | Süsteem peab sisaldama kasutajaõiguste süsteemi, kus juurdepääs koostatud aruannetele ja visuaalidele toimub kasutaja ja kasutajagrupi põhiselt. |
| REQ-95 | 4. Andmeanalüüs | Aruannete trükkimine (pdf) | Süsteem peab võimaldama andmeid väljastada pdf kujul. |
| REQ-96 | 5. Andmete jagamine | Andmete jagamine csv kujul | Süsteem peab võimaldama andmeid väljastada csv failina. |
| REQ-97 | 5. Andmete jagamine | Andmete jagamise veebiteenuste genereerimine | Süsteemis peab olema võimalus ilma süsteemi arendamata genereerida veebiteenuseid andmete jaamiseks. |
| REQ-98 | 5. Andmete jagamine | Andmete jagamine failidena, faili genereerimine | Süsteemis peab olema võimalus ilma süsteemi arendamata genereerida andmetest faile ja need automaatselt paigaldada veebiserverisse. |
| REQ-99 | 5. Andmete jagamine | Visualisatsioonide ja aruannete jagamine veebis | Süsteem peab võimaldama süsteemi haldaja poolt koostatud andmete visualisatsioone (teenuseid) publitseerida veebis sh konkreetsetele nimelistele kasutajatele ilma avaliku juurdepääsuta. |
| REQ-100 | 5. Andmehoidla | Relatsiooniliste andmete salvestamine ja indekseerimine | Süsteem peab võimaldama salvestada andmeid realtsioonilisel kujul. |
| REQ-101 | 6. Andmehoidla | Struktureerimata andmete sh failide salvestamine | Süsteem peab võimaldama salvestada andmefaile ja neid lugeda analüüsipakettide poolt. |
| REQ-102 | 6. Andmehoidla | XML andmete salvestamine ja pärimine | Süsteem peab võimaldama andmeid salvestada xml dokumentidena ning neid salvestatud andmeid kasutada analüüsides. |
| REQ-103 | 6. Andmehoidla | Andmehoidla statistika ja diagnostika | Süsteem peab sisaldada andmehoidla haldamise moodulit, mis võimaldab vaadata kogunenud andmete statistikat ning teha töös olemise ajal diagnostikat koormuse ja jõudluse osas. |
| REQ-104 | 6. Andmehoidla | Hoidla mahuline skaleerimine maksimaalselt kuni 15 minutilise katkestusega | Süsteem peab võimaldama andmebaasi salvestusruumi laiendamist ilma süsteemi tööd katkestamata pikemaks, kui 15 minutit. |
| REQ-105 | 7. Kasutajate autoriseerimine | Kasutaja registreerimine | Kasutaja peab saame end süsteemi kasutajaks registreerida. Registreerumine aktiveerub peale süsteemi administraatori poolset kinnitust. |
| REQ-106 | 8. Kasutajate autentimine | Kasutaja autentimine tunnuse ja parooliga | Kasutaja peab saama süsteemi sisse logida kasutajatunnuse ja parooliga. |
| REQ-107 | 8. Kasutajate autentimine | Kasutajate autentimine ID-kaardiga | Kasutaja peab saama süsteemi sisse logida ID-kaardiga. |
| REQ-108 | 8. Kasutajate autentimine | Kasutajate autentimine Mobiil-ID-ga | Kasutaja peab saama süsteemi sisse logida mobiil-ID-ga. |
| REQ-109 | 8. Kasutajate autentimine | Kasutajate autentimine SmartID-ga | Kasutaja peab saama süsteemi sisse logida SmartID-ga. |
| REQ-110 | 8. Kasutajate autentimine | Parooli meeldetuletus | Süsteem peab sisaldama paroolimeeldetuletuse funktsionaalsust. |
| REQ-111 | 8. Kasutajate autentimine | Kaheastmeline kinnitamine | Süsteem peab sisaldama kahestmelist autentimist nii, et süsteem saadab kasutajale peale selle sisselogimist turvakoodi, mis tuleb sisestada teise aroolina. Kahetasemelist sisselogimist peab saama sisse ja välja lülitada. |
| REQ-112 | 7. Kasutaja autoriseerimine | Kasutajale süsteemi sisenemise loa andmine | Süsteem peab võimaldama kasutajale süsteemi sisenemist lubada ning keelata. |
| REQ-113 | 7. Kasutaja autoriseerimine | Kasutajale objekti ligipääsu õiguse andmine | Süsteemis peab olema võimalik anda kasutajale õigus süsteemis oleva objekti kasutamiseks. Objektideks võivad olla andmetabelid, analüüsipaketid, visualisatsioonid, failid ning süsteemi funktsionaalsused. Funktsionaalsuste lubamine ja keelamise all on silmas peetud andmete alla laadimist, andmete üles laadimist, analüüsipakettide alla ja üles laadimist ning visualisatsioonide alla ja üles laadimist. |
| REQ-114 | 9. Süsteemi monitoorimine | Veebiteenuste monitoorimine | Süsteem peab sisaldama seada üles liidetatud teenuste monitoorimise. Monitoorimise sisu on teenuse töötamise fakti tuvastamine. |
| REQ-115 | 9. Süsteemi monitoorimine | Kasutajaliidese monitoorimine | Süsteem peab sisaldama seada üles süsteemi erinevate osade kasutajaliidese toimimise monitoorimise. Monitoorimise sisu on kasutajaliidese töötamise fakti tuvastamine. |
| REQ-116 | 9. Süsteemi monitoorimine | Kasutaja tegevuste logimine | Süsteem peab logima kõik kasutaja sisselogimised süsteemi. Samuti peab süsteem logima kõik kasutajaõiguste muutmised. |
| REQ-117 | 10. Süsteemi administreerimine | Kasutajate haldus | Süsteem peab sisaldama kasutajate õiguste haldamise lahendust või süsteemi peab saama liidestada LDAP-iga. |
| REQ-118 | 10. Süsteemi administreerimine | Kasutajagruppide haldus | Süsteem peab sisaldama kasutajagruppide haldust või sellise LDAP-igga liidestust, kus on kasutajagruppide haldamine olemas. |
| REQ-119 | 10. Süsteemi administreerimine | Gruppide õiguste haldus (autoriseerimine) | Süsteem peab sisaldama kasutajagruppidele õiguste andmist või liidestust LDAP-iga, kus selline haldus on võimalik. |
| REQ-120 | 10. Süsteemi administreerimine | Serverite haldus | Lahendus peab võimaldama kasutada mitut serverit ning süsteemi konfiguratsioonis peab olema olemas serverite haldamise kasutajaliides, kust on võimalik reaalajas jälgida serverite ressursikasutust ning tekkinud veasituatsioone. |
| REQ-121 | 10. Süsteemi administreerimine | Koormusjaoturite haldus | Süsteem peab võimaldama kasutada koormusjaotureid. |
| REQ-122 | 10. Süsteemi administreerimine | Süsteemi jõudluse tõstmine | Süsteem peab võimaldama süsteemi servereid skaleerida nii automaatselt vastavalt koormusel, kui käsitsi. Skaleerimisel ei tohi süsteemi töö katkeda mitte rohkem, kui 15 minutit. |
| REQ-123 | 10. Süsteemi administreerimine | Süsteemi konfiguratsiooni muudatuste logimine | Süsteem peab logima konfiguratsioonimuudatused, mis on seotud mälu, kõvaketta mahu ning protsessorite arvu muutmisega. |
| REQ-124 | 10. Süsteemi administreerimine | Andmete varukopeerimine | Süsteemi andmed peavad olema varukopeeritavad ja neid peab saama varukoopiast taastada. |
| REQ-125 | 10. Süsteemi administreerimine | Süsteemi konfiguratsiooni varukopeerimine | Süsteemi serverite konfiguratsiooni varukopeerimine peab olema võimalik selliselt, et süsteemi peab saama automaatselt varukoopiast püsti panna. Puhta serveri püsti panemine võib olla lahendatud ka muul automaatsel moel, näiteks konfiguratsioonirepositooriumi abil, millesse kõik konfiguratsioonid salvestatakse. |
| REQ-126 | 10. Süsteemi administreerimine | Haldusliides | Süsteemi konfigureerimiseks peab olemas olema haldusliides. |
| REQ-127 | 10. Süsteemi administreerimine | Failide verisoonihaldus koos failide lukustamisega | Süsteem peab sisaldama failide automaatset versioneerimist nii, et muudatuste tegemisel jääb eelmine faili versioon alles. Faili peab saamamuudatuste tegemise ajaks lukustada. |
| REQ-128 | 11. Masinõpe | Masinõppe algoritmide kasutamine | Süsteemis peab sisaldama masinõppe mootorit, kus on võimalik kasutada masinõppe algoritme ning seadistada masinõppe töövooge. Masinõppe mootor võib olla iseseisev moodul, mis ei ole otseselt seotud analüütika mooduliga. |

**Tabel 7. Suurandmete süsteemi funktsionaalsed ja mittefunktsionaalsed nõuded**

# Suurandmete süsteemi andmed

Suurandmete süsteem peab sisaldama andmeid, mis vastavad järgmistele kriteeriumitele:

1. Andmed on vajalikud Suurandmete süsteemi teenuste loomiseks ja/või osutamiseks, kuid neid ei ole teistes andmekogudes olemas või nende pärimise teenuseid ei ole või teenused ei ole piisava käideldavusega.
2. Andmed vastavad Suurandmete süsteemi kvaliteedistandardile.
3. Andmete kasutamiseks Suurandmete süsteemis ei ole õiguslikke piiranguid ning andmete kasutamiseks on loodud õiguslikud alused ehk Suurandmete süsteemi andmekogu on asutatud.

Juhul, kui andmete kasutamiseks on veebiteenused olemas ja vajalike teenuste osutamiseks pole vajadust andmeid Suurandmete süsteemi kas süsteemi stabiilsuse või jõudluse tõttu kopeerida, siis antud andmeid Suurandmete süsteemi andmebaasi ei tooda.

## Suurandmete süsteemi poolt kasutatavad andmekogud

Käesoleva analüüsi koostamise hetkel on teada, et Suurandmete süsteemis hakkavad olema järgmised andmestikud:

|  |  |  |
| --- | --- | --- |
| Andmestiku kood | Andmestik | Kommentaar |
| AMA | Agrometeoroloogia andmed | Ametlik andmekogu puudub. Projekti käigus koostati mõõdikute kirjeldus, kust selgub täpsem andmekoosseis. |
| TMA | Taimekahjustajate monitooringu andmed | Hetkel ei ole ametlikult andmekogu nende andmete hoidmiseks loodud. Projekti käigus koostati mõõdikute kirjeldus, kust selgub täpsem andmekoosseis. |
| EMÜ | EMÜ erinevate andmekogude andmed, millel täna puudub haldustarkvara. | Ametlikku andmekogu pole loodud. Projekti käigus koostati mõõdikute kirjeldus, kust selgub täpsem andmekoosseis. |
| PANDA | Eesti põllumuldade agrokeemiliste näitajate digitaalne andmekogu | Hetkel puudub ametlik andmekogu. Projekti käigus koostati mõõdikute kirjeldus, kust selgub täpsem andmekoosseis mullaanalüüside osas. PANDA andmete osas võib kujuneda variandiks ka PMK LIS-iga liidestumine. Hetkel ei ole veel LIS ehk PMK laboriinfosüsteem valmis arendatud (see on arenduses). |
| TA | Teeilmajaamade andmed | Hetkel puudub ametlik andmekogu. |
| Metaandmed sh klassifikaatorid | Kasutatavate andmete kirjeldused. | Metaandmete struktuur on kirjeldatud eraldi peatükis. Metaandmed jagunevad mõõdikute ja dimensioonide kirjeldusteks ning andmemudeliteks. |

**Tabel 8. Suurandmete süsteemi kuuluvad andmekogud**

Suurandmete süsteemi poolt kasutatavad ja vahendatavad, kuid Suurandmete süsteemi andmebaasi mittesalvestatavad andmed on:

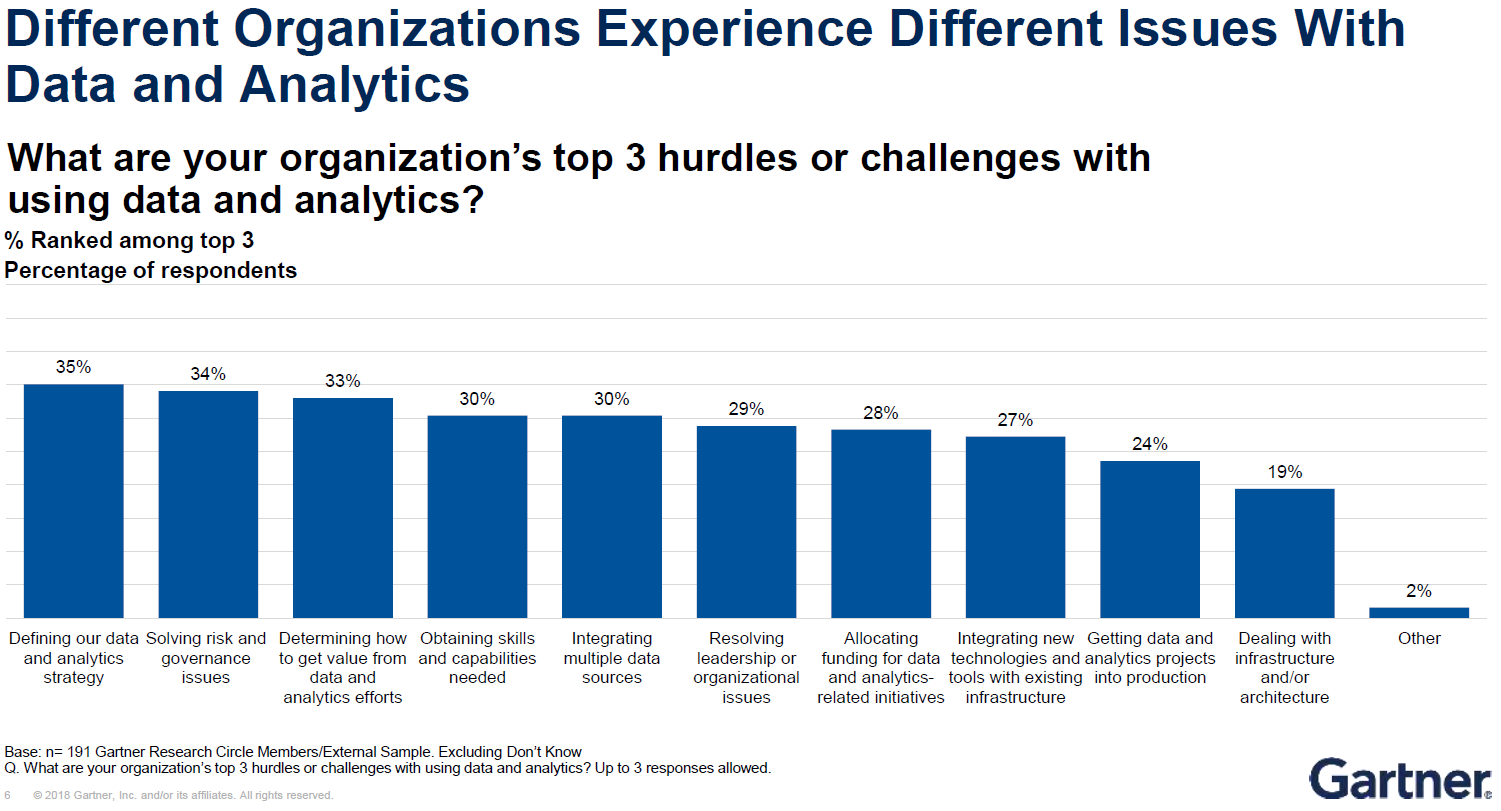
| Andmestiku kood | Andmestik | Kommentaar |
| --- | --- | --- |
| CCS | Nõuetele vastavuse süsteem | Nõuetele vastavuse süsteem sisaldab tootjatele esitatavate nõuete andmeid, mida hetkel jagatakse veebiartiklite vormis. |
| CLIDATA | Riigi Ilmateenistuse poolt kogutud ilmaandmed. | Erinevate allikate ilmaandmetest on plaanis koostada ühtne teenus, mis väljastab kõikide allikate andmeid. See tähendab, et jõudluse kaalutlustel võidakse osasid CLIDATA andmeid kopeerida ka Suurandmete süsteemi andmebaasi. |
| ESTHub | Satelliidipiltide põhjal arvutatud indeksite andmed. | Satelliidipiltide andmeid Suurandmete süsteem ei kasuta. Põllumajanduses läheb tarvis nt rohemassiindeksi andmeid. |
| ETAK | Eesti topograafia andmekogu | Andmeid kasutatakse alusandmetena ruumiandmetega seotud analüüsides ja visualisatsioonides. |
| ETKI | ETKI ja PMK sordivõrdluskatsete andmed. | Olemasoleva sordivõrdluskatsete andmebaasile luuakse veebiteenused, mis võimaldavad andmeid kasutada kolmandates rakendustes. |
| FADN | Põllumajandusliku raamatupidamise andmebaas. | Suurandmete süsteemis luuakse FADN andmete anonüümne visualisatsioonikiht. |
| GEOPORTAAL | Eesti Geoportaal inspire.maaamet.ee | Kasutatakse Geoportaali metaandmeid. |
| KOTKAS | Keskkonnaotsuste infosüsteem | Kasutatakse keskkonnalubade infot. |
| PORTAAL.AGRI.EE | VTA, PMA ja Maaeluministeeriumi portaal, mille kaudu käib teadete ja dokumentide andmevahetus asutuste ning tootjate vahel. | Suurandmete süsteemi kasutaja loomine algab kliendiportaalist, kust kasutaja saab registreeruda Suurandmete süsteemi kasutajaks. Portaalis toimub ka süsteemiga seotud kirjavahetus. |
| TAKS | Toetuste administreerimise ja kontrolli süsteem. | Kasutatakse ja vahendatakse ainult toetuste andmeid. |
| PRIA-TUKS | PRIA turukorraldusmeetmete administreerimise infosüsteem. | Suurandmete süsteem visualiseerib turukorralduse meetmete rakendamise info. |
| MATS | Maaelu arendamise toetuste süsteem. | Vahendatakse ainult toetuste andmeid. |
| MAIT | Maaelu arengukava investeeringutoetuste infosüsteem. | Vahendatakse ainult toetuste andmeid. |
| TSR | Riigi toidu ja sööda käitlejate register. | Vahendatakse käitlejate andmeid. |
| TKV | Taimekaitsevahendite register. | Projekti käigus luuakse veebiteenused, mis võimaldavad andmeid kasutada. |
| MPR | Mahepõllumajanduse register. | Projekti käigus luuakse veebiteenused, mis võimaldavad andmeid kasutada. |
| MATER | Maaparandusalal tegutsevate ettevõtjate register. | Projekti käigus luuakse veebiteenused, mis võimaldavad andmeid kasutada. |
| MSR | Maaparandussüsteemide register. | Projekti käigus luuakse veebiteenused, mis võimaldavad andmeid kasutada. |
| TTR | Taimetervise register. | Projekti käigus luuakse veebiteenused, mis võimaldavad andmeid kasutada. |
| VAETISEREG | Väetiseregister | Projekti käigus luuakse veebiteenused, mis võimaldavad andmeid kasutada. |
| SORDIREG | Sordiregister | Projekti käigus luuakse veebiteenused, mis võimaldavad andmeid kasutada. |
| MIS | Maakataster | Kasutatakse katastriüksuse andmeid. |
| SMKA | Suuremõõtkavaline mullastikukaart koos erinevate mullastiku näitajate andmetega. | Võimalik kasutada koos Maa-ameti XGIS kaardirakendusega. |

**Tabel 9. Suurandmete süsteemi poolt kasutatavad andmekogud**

# Arhitektuurivariandid

Arhitektuur valimise peamised argumendid tulenevad kasutajate vajadustest ja andmete omadustest nagu:

1. andmete keerukus;
2. andmete kogumaht ja ühes teenuses töödeldavate andmete maht;
3. andmemahu kasvu kiirus;
4. andmete ajakohasuse nõuded;
5. rakendusele esitatavad jõudluse nõuded.



**Joonis 6. Analüütika projektides esile kerkivate küsimuste statistika (allikas: Gartner, Inc, 2018)**

Diagrammilt on näha, et suurandmete ja analüütika projektides on peamised küsimused seotud pigem organisatsiooni ja protsesside, kui tehnoloogiatega. Seega ei tasu tehnoloogilisi valikuid üle tähtsustada ega seada omaette eesmärgiks. IT turule on tekkinud hulgaliselt hästi töötavaid tehnoloogiaid, millel puuduvad üksteise ees selged eelised. Pigem on küsimus andmete sisus, nende koosvõimes ehk selles, kas erinevad andmeallikad kasutavad analoogseid andmestruktuure, kas klassifikaatorid on omavahel ühilduvad ja kas andmekvaliteet on piisav ristkasutuse korraldamiseks.

Järgnevas kirjeldame erinevaid aspekte, mille osas on arhitektuuri vaja vaadelda.

## Kogumine vs ühendumine

Üks põhilisi otsustuskohti arhitektuuri valimisel on tasakaalu leidmine andmete kogumise ja andmekogudega ühendumise vahel. Liigne andmete kogumine viib reeglina selleni, et kogutavaid andmeid ei suudeta piisavalt kiiresti uuendada. See viib andmete kvaliteedi langemiseni ajakohasuse ja tõesuse osas. Koguda tuleks ainult neid andmeid, mida teised ei kogu või siis on algallikast andmete kätte saamise võimalused ebapiisava käideldavusega.



**Joonis 7. Tasakaalu otsimine kogumise ja ühendumise vahel (allikas Gartner, Inc, 2018)**

Andmete ristkasutus saab toimuda nii läbi teenuste, kui ka koondades andmed ühtsesse kesksesse andmebaasi. Mõlemal variandil on omad plussid ja miinused.

Plussid andmete kogumise juures:

1. andmed on kiirelt kättesaadavad;
2. andmeid on võimalik kättesaadavaks teha töödelduna ja sobivaimal kujul;
3. andmekvaliteeti on võimalik hinnata kõikselt;

Plussid andmetega ühendumisel:

1. andmed on maksimaalselt ajakohased;
2. liigutatakse ainult neid andmeid, mida on vaja kasutada;
3. ei raisata serveriressursse andmemahule ja töötlusele, mille tulemit ei pruugita tulevikus kasutada;

Kogumise kahjuks räägib peamiselt kogumise aeglus ja suur kulu ning risk, et andmete kogumisel võis minna kogumise protsess kontrolli alt välja ning andmelattu võib tekkida allikast erinev komplekt andmeid.

Ühendumise miinuseks on andmete kasutamise aeglus, kui soovitakse kasutada suurt kogust andmeid. Ühendumise puhul võivad suuremad analüüsid olla väga aeglased, kuna andmeid päritakse allikas üle suhteliselt aeglase ühenduse. Lõppkokkuvõttes võib see tähendada, et analüüsi tulemusi ei saavutata.



**Joonis 8. Ühendumispõhine arhitektuur**

Viimasel ajal on võetud kasutusele termin *Data Hub*, mis sümboliseerib nn andmekeskuste põhist arhitektuuri. Andmekeskus sisaldab endas terve rea sisult lähedaste valdkondade andmebaase ning veebiteenuseid, mis on viidud ühtsete standardite alusel toimivaks.

Eestis on üks *Data Hub*-i näidetest Maa-ameti andmebaasid ja teenused. Need on kirjeldatud ühtses metaandmete süsteemis <http://inspire.maaamet.ee/>.

Samuti võib käsitleda *Data Hub*-ina PMAIS-i, millesse on ühendatud mitmed PMA registrid, kusjuures kõikidel registritel on omad põhimäärused ja õiguslikult on tegemist eraldi andmekogudega.

*Data Hub*-iga toimub andmevahetus reeglina teenuste kaudu, mitte otse andmebaaside vahel. *Data Hub*-i sees võivad olla ühendused mitmesugused, näiteks:

1. REST veebiteenused (näit. MATS).
2. Failivahetus.
3. X-tee teenused

Väljast poolt vaadates peaks *Data Hub* moodustama ühe selgelt piiritletud terviku.

Ühendumine ehk andmete kasutamine vastavalt konkreetsele vajadusele, on elujõulisem variant sel juhul, kui tarvis on kasutada alati ajakohaseid andmeid ja mitte eriti suurtes kogustes. Ühendumist saab kasutada sel juhul, kui andmeallika käideldavuse klass on kõrgem või võrdne andmeid kasutava süsteemi omaga. Vastasel juhul langeb andmetarbija käideldavus. Kui kasutatakse mitut andmeallikat, millel on madal käideldavus, langeb tarbija käideldavusklass veelgi madalamale, kui allikate omad.



**Joonis 9. Kogumispõhine süsteemi arhitektuur (*Data Warehouse*)**

Joonisel on kujutatud tüüpiline kolme tasemega andmeladu, mis andmebaaside mõttes koosneb töötluse andmebaasist, andmelaost ja nn *data mart*-idest ehk analüüsi andmebaasidest.

Kogumise kasuks räägib andmete kiire kättesaadavus. Siiski on see kiirus näiline, kuna andmed on allikas selleks ajaks juba muutunud, kui me neid analüüsi andmebaasist kasutama hakkame.

Andmekogude analüüsi käigus tuvastati, et Eestis on tekkinud mitmeid tehniliselt hästi arenenud andmekeskuseid. Veel hiljuti oli levinud lähenemine selline, et süsteemi väliste sõltuvuste vähendamiseks kopeeriti teise andmekogu andmed enda andmebaasi. Selline lahendus ei tööta aga suurandmete puhul sest paljude andmete osas tuleb muudatusi nii kiiresti peale, et andmete transformeerimise ja kopeerimise ajal tekivad tervikluse probleemid. St kaob järg, kus andmete ülekandega ollakse, kuna ümbritsevad andmed on juba muutunud. Ka ei ole võimalik andmeid ühe tervikliku transaktsiooni sees kopeerida, sest selline päring töötaks väga pikalt ja reeglina lõpeks lihtsalt serveri mälu täis saamisega. Selliseid andmeid, millel on olemas toimiv teenus andmete jagamiseks, ei ole mõtet täies mahus kopeerida kohalikku süsteemi. See ei tähenda seda, et kui liidestatud andmekogu andmeid kasutatakse näiteks mõnes dokumendis, siis sinna dokumenti ei tohiks andmeid kopeerida. Seda tuleb ikkagi teha, kui on oluline säilitada andmete hetkeseis, et hiljem oleks võimalik kindlaks teha, milliste andmete pealt otsuseid tehti.

Nii Suurandmete süsteemi analüüsil, kui ka mujal maailmas on jõutud järeldusele, et mastaapsete andmeladude moodustamine ei pruugi olla toimiv lahendus. Andmete maht, struktuuri keerukus ja muutumise kiirus ei võimalda reaalajas toimivat andmeladu luua. Seega tuleks leida tasakaal andmetega ühendumise ja andmete kogumise vahel. Koguda tuleks ainult neid andmeid, mida ei ole võimalik andmekogudest reaalaajas kätte saada. Erijuht on siin ka mahukate analüüside tegemine, kus on vaja kogu andmestikku korraga ja/või korduvalt kasutada. Viimasel juhul on kogu andmekomplekti ühekordne kopeerimine analüüsikeskkonda õigustatud.

Seega ka loodava Suurandmete süsteemi loomise õnnestumise eelduseks on tasakaalu leidmine andmete kogumise ja andmetega ühendumise vahel. Vältimatu on arvestada juba tekkinud andmekeskustega nagu ESTHub, Inspire, PMAIS ning PRIA toetuste ja põllumassiivide register. Nende andmeid ei ole otstarbekas püsivalt ja täies mahus dubleerida.

Osadest andmetest on andmelao moodustamine Suurandmete süsteemi puhul vältimatu. Nendeks on sellised põllumajandusvaldkonna andmed, millel täna puudub andmekogu, kuid mis on olulised. Need andmed tuleks seadustada, kui Suurandmete süsteemi andmed.

## Pilv vs lokaalne paigaldus

On olemas kaks põhimõtteliselt erinevat süsteemi paigaldamise viisi:

1. paigaldus pilve;
2. lokaalne paigaldus.

Kolmandaks variandiks võib lugeda ka privaatpilve, mis ühendab endas osad pilve ja lokaalse paigalduse eelised.

| Omadus | Pilv | Lokaalne | Kommentaar |
| --- | --- | --- | --- |
| Paigalduse kiirus | + | - | Paigalduse kiirus sõltub sellest kuivõrd on paigaldus automatiseeritud. Kuna pilves korratakse pidavalt paigaldusi, siis on paigaldusprotsess paremini välja arendatud ja ka automatiseeritav. |
| Paigalduse maksumus | + | - | Pilves virtuaalserverite paigalduse tasu ei ole. Samuti ei tule osta litsentse ega riistvara. Tasuda tuleb hiljem teenuste eest. |
| Keskkonna stabiilsus | + | - | Pilve stabiilsus infrastruktuuri tasemel on reeglina parem, kui lokaalsel paigaldusel, sest kasutatakse virtuaalmasinaid. |
| Turvalisus | + | - | Turvalisus on pilves selle tõttu parem, et ründeid monitooritakse mastaapsemalt ning turvasüsteemid on paremini välja ehitatud. |
| Monitoorimine | + | - | Tuntumates ja suuremates pilvedes on välja arendatud väga põhjalikud monitoorimise lahendused nii teenuste tervise, kui ka ressursikasutuse osas. |
| Ressursikulu monitooring | + | - | Pilve puhul on arveldamise aluseks kasutatud serveriressurss. Selle tõttu tuuakse pilveteenuste arvetel ressursikasutus välja äärmiselt täpselt ja paljude detailidega. |
| Maksumus maksimaalse ressursikasutuse juures | - | + | Lokaalse paigalduse eeliseks on parem rahaline efektiivsus, kuna serverite ressursi täielikul kasutamisel kasvavad kulud suhteliselt vähe. Pilves sõltub teenuse maksumus kõige rohkem sellest, kui suurt pinda ja mitut protsessori tuuma on kasutatud. |
| Üles ja alla skaleerimise lihtsus | + | - | Süsteemi kõvaketta ja protsessorite laiendamine või vähendamine käib pilves hiireklikiga või sootuks automaatselt. Lokaalse paigalduse puhul võib see olla keerulisem. |
| Skaleerimise maksumus | + | - | Skaleerimine ise ei makse pilves midagi. Maksta tuleb ainult ressursi kasutuse eest. Lokaalse paigalduse korral tuleb teha laiendava skaleerimise korral investeering kohe. Eriti efektiivne on pilves skaleerimine pidevalt muutuva koormuse korral. Kuna arveid esitatakse detailse infoga, siis on võimalik ka saadud kuluarve jagada tarbijate vahel tuues välja konkreetsed maksumuse arvutamise alused. |
| Halduskulud | + | - | Pilves väikese koormusega süsteemi ülal hoidmise eest on arve pea olematu. Maksab ainult ressursi tarbimine. |
| Kontroll andmete üle | - | + | Lokaalses paigalduses on kontroll andmete üle parem, kui pilves. Siiski on lokaalses keskkonnas riskid probleemide tekkimiseks suuremad, kuna on reeglina rakendatud vähem turvameetmeid. |
| Paigalduse abistaja ja funktsionaalsuse soovitaja olemasolu | + | - | Laiemalt levinud Azure ja Google pilvedes on olemas ekpertsüsteem, mis abistab serveri haldajat pidevalt. Sh süsteemi paigalduse turvalisuse küsimustes. |

**Tabel 10. Pilve ja lokaalse paigalduse eelised ja puudused**

Juhul, kui privaatpilve infrastruktuur ja haldusmeeskond on hästi välja arendatud, siis ei jää privaatpilv ehk kohalikule infrastruktuurile paigaldatud pilv oluliselt alla suurtele kommertspilvedele. Olemas on ka variant, et süsteemi tööd alustades kasutada kommertspilve, vältimaks suurt investeeringut teadmata, kui kiiresti loodav süsteem rakendub. Alles kriitilisema ressursikasutuse tekkimisel tuleks kaaluda süsteemi kolimist privaatpilve, et vähendada ressursi kasutamise kulusid.

## Toega vs toeta baastarkvara

Kasutataval baastarkvara ei saa rakendada ilma toeta. Tugi väljendub pidevas tarkvara funktsionaalsuse täiendamises ja turvaaukude paikamises. Riiklikud infosüsteemid ei saa turvalisuse kaalutlusel töötada ilma toeta. See tähendab, et tugi tuleb kas endal organiseerida, mis tagaks baastarkvara pideva täiendamise või kasutada toega tarkvara. Üheks toe tagamise võimaluseks on suure kasutuajate kogukonnaga (*community*) tarkvara kasutamine, mis lihtsustab toe tagamist.

## Suletud lähtekood vs avatud lähtekood

Täielikult suletud lähtekoodiga baastarkvara kasutamine ei ole soovitav kasutada. Avatud lähtekood peab olema olemas vähemalt liideste ja adapterite osas, et andmete liikumise juures probleemide tekkimisel oleks võimalik kiiresti sekkuda ja liidest parandada või täiendada nii, et süsteem ei muutuks blokeerituks pikemaks ajaks.

## Andmevahetus andmekogudega – veebiteenused, andmebaasiühendused, failid

Riigiinfosüsteem koosneb suurest hulgast andmekogudest, mis suuremal või vähemal määral on kirjeldatud riha.ee portaalis. Käesoleva projekti käigus vaadeldud riiklike andmekogude arv on 41. Riiklikuks andmevahetuse kanaliks on X-tee. X-tee sobib juhul, kui põhiliselt vahetatakse andmeid väikeses mahus kas ühe isiku või ettevõtte kohta. X-teel muutub andmevahetus praktilises kasutuses ebaefektiivseks, kui tuleb vahetada suures mahus andmeid ja seda reaalajas.

Avaandmete osas on riiklike andmekogude osas mõistlik võtta kasutusele RESTful veebiteenused, täpsemalt Open Data standardil põhinevad teenused, mille rakendamisest on täpsemalt räägitud käesoleva dokumendi standardite peatükis. Hetkel on Eestis väga vähe neid andmekogusid, kus sellised teenused (RESTful) juba loodud on. Eeskujulikeks näideteks on Maa-ameti ja Statistikaameti osad andmestikud.

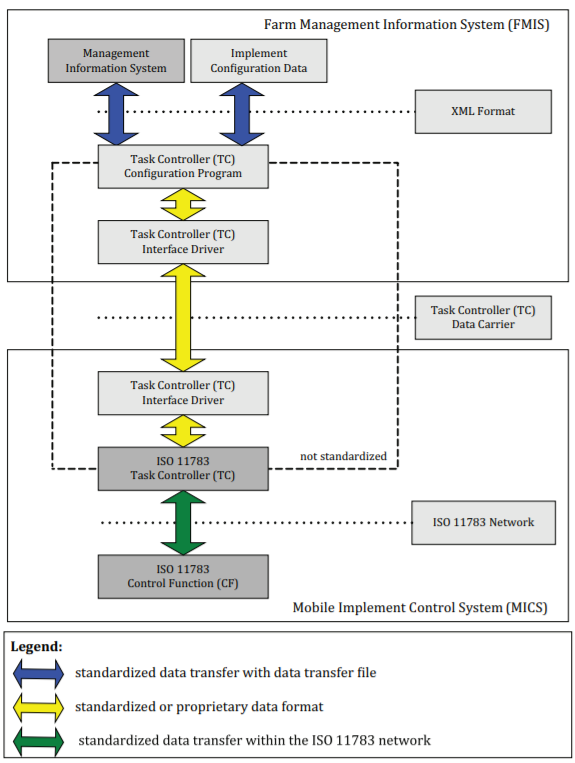
Käesolevas on oluline märkida, et andmete jagamise osas saame tulevikus kasutada järgmisi variante:

1. Iga andmekogu loob endale andmejagamise teenused ning Suurandmete süsteem pakub nende pealt oma sünteesimise teenuseid.
2. Suurandmete süsteem hakkab pakkuma nii andmejagamise teenust, kui ka sünteesimise teenuseid.
3. Andmejagamise teenust hakkab pakkuma Statistikaamet, mis tegutseb vastavalt uuele Statistikaseadusele. Suurandmete süsteem saab andmed Statistikaameti kaudu ning pakub omi teenuseid sealt saadud andmete pealt.

Tõenäoliselt võib siin olla kesktee kõige parem variant. See tähendab, et Statistikaamet võiks hakata pakkuma andmete jagamise teenust nende andmete osas, mida Suurandmete süsteemi teenused täna veel ei kasuta ning Suurandmete süsteem kasutab konkreetse andmekogu juurde rajatud teenuseid, et pakkuda enda lisaväärtust tootvaid teenuseid. Tuleb teha strateegiline otsus, kas Suurandmete süsteem hakkab jagama ainult enda poolt rikastatud andmeid või ka erinevate andmekogude üksikandmeid ilma neid rikastamata ja sünteesitud andmeid juurde lisamata. Majandusliku otstarbekuse aspektist lähtudes ei oleks soovitav, et Suurandmete süsteem hakkab pakkuma samas vormis ja andmekoosseisus andmejagamise teenust, mis Statistikaametki. Suurandmete süsteem võiks pakkuda ise teenust juhul, kui andmete jagamise vorm või sisu ei ole sama, mis Statistikaameti andmejagamise teenusel.

## Andmevahetus põllumajandusmasinatega

Analüüsi tulemusel selgus, et tootjate jaoks on arvestatavaks probleemiks andmevahetus põllutöömasinatega. Erinevad masinatootjad on välja arendanud omad süsteemid. Need süsteemid ei ole omavahel ühilduvad. Ka on reeglina antud süsteemid tasulised, kui soovida täisfunktsionaalsust kasutada. See tähendab, et kui tootjal on n erinevat marki masinat, siis ta peab üldjuhul kasutama ka n tarkvara ning maksma n lepingutasu. Üldjuhul ei ole võimalik kasutada sama juhtpulti erinevate tootjate masinatel, sest need on tihedalt seotud masinatootja poolt loodud tarkvaraga.



**Joonis 10. Standardil ISO 11783 põhinev andmevahetuse skeem (allikas ISO)**

Joonisel toodud skeem iseloomustab masina ja farmi infosüsteemi (FIM) andmevahetust. Andmevahetus on kahesuunaline. Masinale või seadmele antakse farmi süsteemist ette juhtandmed, näiteks väetamise kaart. Ette antud kaardi ja GPS signaali abil juhib masin erinevaid agregaate, näiteks sulgeb ja avab düüse väetise või taimekaitsevahendi pihustamiseks.



**Joonis 11. Põllumajandusmasina andmevahetusrakenduse liidese arhitektuur**

Eelpool kirjeldatust tingituna tehti järeldus, et Suurandmete projekti käigus tuleb luua andmevahetuse tarkvara masinatega, mis tagaks kõikide masinate osas, mis vastavad ISO 11783 standardiperekonna nõuetele, andmevahetuse ühesugusel kujul. Seda eelkõige masina töö andmete lugemisel ja ePõlluraamatusse salvestamisel. Andmete kasutamisel tuleb arvestada, et nende kvaliteet võib olla erinevate masinate puhul erinev. Enne ePõlluraamatusse andmete salvestamist tuleb andmed vastavalt vajadusele puhastada. Isegi, kui andmeid tuleks puhastada, tähendaks automaatandmevahetus masinaga arvestatavat ajalist võitu põlluraamatu täitmisel.

Samuti tuleb arvestada, et igal masinatootjal on omad reeglid (masina kasutajaleping), millistes piirides on andmete kasutamine lubatud. Ühest küljest on piirangud seotud masina ohutuse ja töökindlusega ning teisest küljest ka masinatootja autoriõigustega.

ISO standard kirjeldab konkreetse andmestruktuuri, mille kaudu andmeid vahetatakse. Järgnevas on toodud projekti raames tehtud traktori testsõidu tööandmed ISO standardis kirjeldatud kujul:

<?xml version="1.0" encoding="UTF-8"?>

<ISO11783\_TaskData VersionMajor="2" VersionMinor="0" ManagementSoftwareManufacturer="AgLeader Technology" ManagementSoftwareVersion="4.0\_7537" TaskControllerManufacturer="AgLeader Technology" TaskControllerVersion="4.0\_7537" DataTransferOrigin="2">

<TSK A="TSK-0" B="2018-12-10\_12:08:43" G="4" C="CTR-0" D="FRM-0" E="PFD-0">

<TIM D="4" B="2018-12-10T12:27:13">

<DLV A="0051" B="896" C="DET-1"/>

<DLV A="0051" B="896" C="DET-1"/>

<DLV A="0051" B="896" C="DET-1"/>

<DLV A="0051" B="896" C="DET-1"/>

<DLV A="0051" B="896" C="DET-1"/>

</TIM>

<DAN A="A00A80000B202AB4" C="DVC-0"/>

<PAN A="PDT-0" E="DET-1">

<ASP D="4" A="2018-12-10T12:09:58"/>

</PAN>

<PAN A="PDT-0" E="DET-1">

<ASP D="4" A="2018-12-10T12:12:13"/>

</PAN>

<PAN A="PDT-0" E="DET-1">

<ASP D="4" A="2018-12-10T12:17:13"/>

</PAN>

<PAN A="PDT-0" E="DET-1">

<ASP D="4" A="2018-12-10T12:22:13"/>

</PAN>

<PAN A="PDT-0" E="DET-1">

<ASP D="4" A="2018-12-10T12:27:13"/>

</PAN>

<TLG A="TLG00000" C="1"/>

</TSK>

<PFD A="PFD-0" C="KOMMU" D="132277" E="CTR-0" F="FRM-0">

<PLN A="1">

<LSG A="1">

<PNT A="2" C="58.8034226" D="26.7522538"/>

<PNT A="2" C="58.8033372" D="26.7523559"/>

<PNT A="2" C="58.8031837" D="26.75251"/>

<PNT A="2" C="58.803341" D="26.7533593"/>

<PNT A="2" C="58.8033448" D="26.753394"/>

<PNT A="2" C="58.803337" D="26.7534106"/>

<PNT A="2" C="58.8033237" D="26.7534067"/>

</LSG>

<LSG A="2">

<PNT A="2" C="58.8035278" D="26.7559649"/>

<PNT A="2" C="58.8035322" D="26.7560398"/>

<PNT A="2" C="58.8035173" D="26.7561103"/>

<PNT A="2" C="58.8034909" D="26.7561598"/>

<PNT A="2" C="58.8034562" D="26.7561828"/>

</LSG>

<LSG A="2">

<PNT A="2" C="58.804132" D="26.7551101"/>

<PNT A="2" C="58.804149" D="26.7552204"/>

<PNT A="2" C="58.8041426" D="26.7553087"/>

</LSG>

</PLN>

</PFD>

<FRM A="FRM-0" B="TO" I="CTR-0"/>

<CTR A="CTR-0" B="KU"/>

<DVC A="DVC-0" B="New Holland T7.2, Kverneland EDW" D="A00A80000B202AB4" F="00000000000000" G="FF000000000000">

<DET A="DET-0" B="1" C="1" E="0" F="0"/>

<DET A="DET-1" B="2" C="2" E="1" F="1">

<DOR A="3"/>

<DOR A="4"/>

</DET>

<DET A="DET-2" B="5" C="4" E="2" F="2">

<DOR A="6"/>

<DOR A="7"/>

</DET>

<DET A="DET-3" B="8" C="4" E="3" F="2">

<DOR A="9"/>

<DOR A="10"/>

</DET>

<DET A="DET-4" B="11" C="4" E="4" F="2">

<DOR A="12"/>

<DOR A="13"/>

</DET>

<DPT A="3" B="00B3" C="1"/>

<DPT A="4" B="0043" C="24000"/>

<DPT A="6" B="0087" C="-11000"/>

<DPT A="7" B="0043" C="2000"/>

<DPT A="9" B="0087" C="-9000"/>

<DPT A="10" B="0043" C="2000"/>

<DPT A="12" B="0087" C="-7000"/>

<DPT A="13" B="0043" C="2000"/>

<DPT A="15" B="0087" C="-5000"/>

<DPT A="16" B="0043" C="2000"/>

<DPT A="18" B="0087" C="-3000"/>

<DPT A="19" B="0043" C="2000"/>

</DVC>

<PDT A="PDT-0" B="Amm. Nitrate"/>

</ISO11783\_TaskData>

Täpsem info ISO standardi ja selles kasutatavate erinevate näitajate kohta on toodud veebilehel <https://www.isobus.net/>.

Selleks, et ISO põhiseid andmeid kasutada, tuleb farmi infosüsteem ja ka riiklik ePõlluraamat viia vastavusse ISO 11783 standardis kirjeldatud nõuetega.

Alternatiivseteks variantideks masinaandmete vahetamisel on:

1. Luua riiklikul tasemel andmevahetusmoodul, mis CAN BUS tasemel andmeid loeb.
2. Luua adapterid kõikide kasutatavate masinate tootjate tarkvaradega.

Mõlemal variantidel on ka miinuseid. Esimese puhul on õigete andmete kätte saamine ja töötlus küllalt keerukas. Teise variandi miinuseks on suur osapoolte arv ja õiguslikud takistused andmete kasutamiseks.

Kaasnevad tegevused mõlema variandi realiseerimisel on osalemine EL tasemel kompetensiliidrina loodud avatud tarkvara näitel ning näidata selle kasu ühiskonnale ning saavutada olukord, kus masinatootjate andmed soovitud andmekoosseisus vabastakse kasutamiseks kindlaks määratud eesmärkidel (nt andmekorje automatiseerimiseks). Vabastamise all peame eelkõige silmas masinate kasutajalepingute muutmist mõistlikumaks tootjapoolse andmete kasutamise õiguste osas. Käesoleva projekti käigus koostati nimekiri andmelementidest (mõõdikutest), milliseid andmeid masinatest vaja oleks.

EL tasemel on vaja luua motivatsioon (nt positiivne soodustususena või negatiivne piirangutena) masinatootjale, et nad oleksid andmete kasutamise õiguse andmisest huvitatud.

Õiguslikeks instrumentideks masinaandmete vahetamise standardiseerimiseks on kehtestada Euroopa Liidus direktiivi abil ISO 11783 kasutamise kohustuslikkus täpsustades ühtlasi kasutatavate andmete koosseisu, mis peab olema iga masinatootja masinatest standardsel kujul kättesaadav. Andmekoosseisu aluseks on võimalik võtta eelpool nimetatud nimekiri.

Pilootprojekt masinaandmete hõivetarkvara loomise osas tooks potentsiaalselt esile ühiskondliku kasu ning seeläbi toetuse jõulisemate sammude astumiseks. Riskiks on tootjate mitte kaasatulemine loodava motivatsiooniga, kuid võimaluseks see, et keegi on esimene ja annab andmete kasutamise õiguse, misjärel konkurents sunnib ka teisi järgnema. Oluline on mentaliteedi muutus ja põllumajandustootja õiguste austamine tema enda andmete kasutamisel. Siinkohal tuleb jälgida NIVA projekti edusamme antud valdkonnas, milles osaleb Eesti poolelt PRIA, masinandmete teemat veab projekti Hollandi partner.

# Tehnilise arhitektuuri variandid

Tehnilise arhitektuuri osas pakume välja 3 variant:

1. SQL Server 2019 Polybase ning Azurel ja Power BI-l põhinev arhitektuur.
2. Apache Hadoopil ja MapReducel põhinev arhitektuur.
3. Sparkil ja PostGIS-il põhinev arhitektuur.

## Andmebaaside jõudlusest

Kui rääkida relatsioonilistest andmebaasidest, siis andmemahu osas on olulised aspektid kirjete arv ja baitide arv kirjes.

Järgnevas tabelis on näide kolme ruumiinformatsiooni töötlemise võimekusega andmebaasisüsteemi jõudlusnäitajate kohta.

|  |  |  |  |
| --- | --- | --- | --- |
| Maht (punkti) | PostGIS (ms) | Spatial Spark (ms) | Elcano (ms) |
| 1000 | 234 | 6 543 | 9 516 |
| 10 000 | 326 | 6 622 | 9 714 |
| 100 000 | 3783 | 8 301 | 9 030 |
| 1 000 000 | 29 899 | 8 301 | 10 747 |
| 10 000 000 | 269 257 | 20 487 | 17 099 |
| 100 000 000 | 5 752 821 | 55 017 | 37 378 |
| 1 000 000 000 | Rohkem, kui 10h | 399 100 | 273 074 |

**Tabel 11. Erinevate andmetöötlussüsteemide jõudlus andmete töötlemisel (allikas [1])**

Jõudlustestist on näha, et süsteemide võimekusel on suur vahe. Kas see vahe on oluline Suurandmete süsteemi kasutajale? Kui vaadata Suurandmete süsteemi teenuste loetelu, siis esialgu välja tulnud vajadused on peamiselt seotud põllu dimensiooniga. Näiteks põllu huumusbilansi arvutamise teenus. Seda teenust kasutades töödeldakse maksimaalselt ehk 10 000 andmekirjet. Sellise kirjete arvu juures on spetsiaalselt suurandmete töötlemiseks mõeldud tehnilised lahendused väga aeglased, ületades päringu vastusega üldtunnustatud kriitilise ooteaja, mis on ca 6-7 sekundit. Seega põllupõhiste teenuste loomisel oleks ekslik kasutada suurandmete tehnoloogiat.

Suurandmete töötlemise tehnoloogiat on mõistlik kasutada siis, kui mahukate (antud näite puhul 1 000 000 000 kirjega) andmete analüüse tehakse rohkem, kui umbes 120 korda päevas. Hetkel ühtki sellist laiatarbe teenust teada ei ole, kus nii palju andmeid töödeldakse. Seega on suurandmete tehnoloogia rakendamine üldkasutatavate teenuste osas küsitava väärtusega.

Jõudlustestide tulemused näitavad, et hetkel teada olevate teenuste ressursivajadus ei ületa tavalist andmekogu tarkvara ressursivajadust. Seega sobib andmete töötlemiseks andmebaasi osas üldjuhul Postgre ja PostGIS tarkvara.

Küll aga võib suurandmete töötlemise tehnoloogiat vaja minna teenuste ettevalmistustöödel. Seega teadusasutustel sel juhul, kui jooksutatakse analüüse kogu Eesti andmestikuga. Ühe kaardikihi maht LIDAR-i poolt mõõdistatuna võib ulatuda 45 227 000 000 punktini, kui mõõdistamise detailsus on 1x1 meetrise punktiga. Nii suurte andmemahtude korral ei ole tavaline Postgre andmebaas enam reaalajas kasutatav. Sellise mahu töötlemiseks on tarvis eraldi teistsuguse arhitektuuriga keskkonda.

## Lahenduse horisontaalne skaleeruvus

Lahenduse skaleerimiseks on olemas erinevaid variante. Pilveteenuse kasutamisel on küllalt lihtne tõsta süsteemi jõudlust kasvatades protsessorite arvu, samuti kasutatavat kettamälu. Kui siin on piir saavutatud on järgmiseks sammuks lahenduse tükeldamine. Kõikide pakutavate arhitektuurivariantide osas on võimalik luua teenused erinevate rakendustena olgu see andmete visualisatsioon või veebiteenus arvutuste tegemiseks. Need rakendused ei pea jooksma ühel serveril ega isegi ühes pilves. Ära on võimalik kasutada tasuta pilveteenused (näit. Power BI) ning ressursi vajaduse kasvades minna üle teistele.

Kasutajate arvu kasvades hakkab Azure ja Google pilve kulud kasvama ja selle juhtudes on võimalik tuua enim kasutatavad rakendused privaatpilve, kui infrastruktuuri haldamise võimakus on kohapeal olemas.

Süsteemi enim kasutatav andmestruktuur näitajate ja sündmuste osas on planeeritud selline, et väga lihtne on transformeerida andmed mitmeks andmebaasiks. Samuti üheks kokku tagasi, kui tarvis. Planeeritud integratsioonivahendites ei ole vahet, kas andmed tulevad ühest või mitmest andmebaasist. Seega on võimalik ilma keerulisemaid tehnoloogiaid kasutamata süsteemi päringud jaotada n serveri peale. Antud juhul on n sündmuseliikide arv.

Ainsana tuleks ühele andmebaasiserverile planeerida metaandmete haldus, mis juhib kogu lahenduse toimimist.

## Arhitektuuri variant nr 1

Microsoft on välja arendanud juhtiva andmete halduse ja analüütika perekonna, mis koosneb:

1. **MS SQL Server 2019 Polybase** – tarkvara, mis võimaldab läbi ühtse SQL keele ja ODBC ning JDBC liideste kasutada nii SQL Serveri enda, kui ka muude andmebaaside sh HDFS (Hadoop File System), Oracle, Teradata ja Azure Data Lake Store andmeid. Andmebaasil on olemas GIS andmete funktsionaalsus nii andmetüüpide (*geometry, geography*), kui ka laialdase funktsionaalsuse näol. SQL Serveri funktsionaalsusesse kuulub analüütiliste rakenduste osas eriti oluline funktsionaalsus – *in-memory* OLTP transaktsioonid ja *columnstore* indeksid, mis võimaldavad päringute jõudlust kümnekordistada. Tehnoloogia põhineb andmete operatiivmälus hoidmisel ja mälust kiirel leidmisel. SQL Serveri puhul on võimalik kasutada klasterdamist (*Always On Failover Cluster*), mis tähendab katkestuskindlat lahendust, kus klastrisse on ühendatud üksteist dubleerivad serverid.
2. **SSMS** – SQL Server Management Studio, mis võimalda hallata SQL Server andmebaasi, mudeleid ja skripte andmebaasi loomiseks ning muutmiseks.
3. **SQL Server Analysis Services** – andmeanalüütika mootor, mis sisaldab erinevaid tarkvaratooteid aruandluse ja analüüside tegemiseks.
4. **SSIS** – SQL Server Integration Services on andmete integreerimise ja transformeerimise platvorm, millel on olemas ühendumisvõimalused nii failide (csv, xlsx), kui erinevate relatsiooniliste andmebaasidega.
5. **Power BI Pro**– võimas andmete integreerimise, analüüsi ja analüütiliste rakenduste koostamise vahend koos tasuta pilveteenusega, mis võimaldab andmeid integreerida nii relatsioonilistest andmebaasidest Postgre, Oracle, SQL Server jt. ning samuti veebiteenustest, Sparki API-st ja HDFS-ist. Power BI-l on sisse ehitatud ka R-i ja Pythoni skriptimise tugi, mis võimaldab teha keerukamaid andmeanalüüse ja töötlusi ilma täiendavaid keskkondi püsti panemata. R võib paikneda ka eraldi serveril, kui on tarvis suuremat jõudlust.
6. **Sharepoint** – dokumendi ja listide haldus, versioonihaldus.
7. **Teams** – grupitöö dokumentidega (xlss, docx, pptx).
8. **Azure** – Microsofti poolt hallatav pilveteenus. Teenus võimaldab kasutada nii Postgres, kui SQL Serveri andmebaase, rakendusservereid Java ja C# rakendustele ning pakub kogu infrastruktuuri koormuse jagamise, monitoorimise ja turvamise võimalusi väga laial skaalal.
9. **Windows Server** – operatsioonisüsteem.

Power BI-l, SQL Serveril ning Azure pilveteenusel põhineb ka projekti käigus koostatud andmekogude analüüsi tulem: tietoanalytics.ee/PRIA.

Lahenduse plussideks on:

1. Toimiv kahetasemeline autoriseerimise ja autentimise süsteem, mis on võimalik ühendada kohaliku domeeniga (*Active Directory’*ga). Samad kasutajad nii pilves, kui kohalikes paigaldustes.
2. Automaatne skaleerumine (kui pilves sisse lülitatud).
3. Pilvepaigalduse puhul käsitsi skaleerimine hiireklikiga.
4. Tugev ja läbi proovitud sama perekonna erinevate vahendite integratsioon.
5. Lihtne toodete paigaldus, kuna tarkvara on paketeeritud paigalduskomplektidesse.
6. Paketid on toetatud pilveteenuses.
7. Pidev ja kiire tarkvaraareng. Automaatuuendused nii serverites, kui töökohatarkvarades, koos teavitussüsteemiga.
8. Andmeadaterid on Power BI-l avatud lähtekoodiga.
9. Nii adapterteenused, kui visualiseerimise komponendid on arendatavad kasutaja poolt.
10. On olemas nii ruumi- kui tärkandmete visualiseerimise võimalused.
11. Power BI REST API võimaldab koostatud visualisatsioonide andmeid kasutada ka teistel süsteemidel.
12. Spetsialistide olemasolu Eestis.

Miinused:

1. Suletud lähtekood baastarkvaral.
2. Osadel tarkvaradel tasulised litsentsid (SQL Server ja Windows Server).
3. Tarkvara kiire areng sunnib arendama ka platvormile rajatud teenuseid.

## Arhitektuuri variant nr 2

Üks arhitektuuri variantidest on kasutada Apache projektides välja arendatut:

1. **Apache Hadoop** – avatud lähtekoodiga teek, mis on välja arendatud suuremahuliste andmete haldamiseks ja analüüsimiseks võimaldamaks tööd tuhandetel arvutitel korraga.
2. **HDFS** – Hadoop Distributed File System on failisüsteem, mis võimaldab andmete klasterdamist ja pakub võimalus väga kiireks andmete pärimiseks sh agregeerimiseks.
3. **Apache Yarn** – arvutusvõimsuse haldamise ja monitoorimise rakendus.
4. **Appache MapReduce** – Võimaldab luua andmetöötluse (paralleel) protsesse, mis ainult Hadoopi vahenditega ei ole realiseeritavad.
5. **Appache Zeppelin** – andmete visualiseerimise rakendus.
6. **Linux** – operatsioonisüsteem.

Plussid:

1. Avatud lähtekood.
2. Tasuta litsentsid.
3. Ühtne süsteem võib töötada nii ühel, kui ka tuhandetel masinatel üheaegselt.
4. Tõrgete automaatne käsitlemine. Vigane riistvara on võimalik asendada ilma süsteemi tööd katkestamata.
5. Süsteemi horisontaalne skaleerimine ilma katkestuseta.
6. Zeppelin REST API võimaldab koostatud visualisatsioonide andmeid kasutada ka teistel süsteemidel.

Miinused:

1. Tasuta versioonidel puudub ametlik tugi ja see tuleb endal organiseerida.
2. Puudub tarkvarauuenduste automaatse levitamise süsteem.
3. MapReduce ei võimalda *in-memory* andmete töötlust, mis teeb selle märkimisväärselt aeglasemaks paljudest teistes paralleeltöötluse lahendustest.
4. Enamus tööst käib käsureal. Pole välja arendatud graafilist kasutajaliidest, kus programmeerimisoskusteta kasutaja saaks andmeid analüüsida.
5. Tasuta versioonide osas pole väljaarenenud paketeeritud paigalduskomplektide süsteem, mis teeb vajalike komponentide otsimise ja leidmise keeruliseks ning väga ajamahukaks.
6. Vaikimisi ei ole olemas võimalus autentimise ja autoriseerimise ühendamiseks kohaliku LDAP-iga.
7. Hadoop on tänaseks juba vana tehnoloogia ning hakkab võimaluste poolest alla jääma uutele pilves pakutavatele suurandmete käsitlemise toodetele. Sisuliselt on pilved realiseerinud vaikimisi Hadoopi eelised.
8. Eestis puudub piisav spetsialistide võrgustik, kes oleks võimeline süsteemi paigaldama ja haldama.

## Arhitektuuri variant nr 3

Alternatiivne arhitektuur Hadoopile ja MapReducele:

1. **Apache Spark** – paralleeltöötluse raamistik, mis võimaldab teha *in-memory* töötlusi ja analüüse. Kuna nii *in-memory* andmekasutus ja paralleeltöötlus on üheaegselt võimalikud, siis on jõudluse näitaja märkimisväärselt kõrged.
2. **GeoSpark** – Sparki laiendus ruumiinfo analüüside tegemiseks.
3. **GeoSparkViz** – Sparki laiendus, mis võimaldab ruumiandmeid visualiseerida.
4. **Appache Zeppelin** – andmete visualiseerimise rakendus.
5. **Postgre** – Maaeluministeeriumi haldusalas laialt levinud andmebaasimootor, mille kasutamise tuleks jätkuvalt kaaluda ka Suurandmete süsteemi relatsiooniliste andmete hoidmiseks ja töötlemiseks.
6. **PostGIS** – laiendus Postgre andmebaasil, mis võimaldab teha operatsioone ruumiandmetega. PosGIS ei sisalda automaatset klasterdamist. Klastrid tuleb luua ise jagades andmed sellisteks alamandmebaasideks, kus asub koos korraga töötlust nõudev andmestik. Selliselt on võimalik efektiivselt ära kasutada *in-memory* võimalused Sparki puhul.
7. **MongoDB** – andmete salvestamiseks XML ja JSON dokumentide kujul.
8. **Pgadmin** – Postgre andmebaasi haldamise tarkvara.
9. **Linux** – operatsioonisüsteem.

Plussid:

1. Avatud lähtekood.
2. Tasuta litsentsid.
3. Ülikiire andmete töötlemine ja analüüs.
4. On olemas ruumiinfo kui tärkinfo visualiseerimise võimalused.
5. Zeppelin REST API võimaldab koostatud visualisatsioonide andmeid kasutada ka teistel süsteemidel.

Miinused:

1. Tasuta versioonidel puudub ametlik tugi ja see tuleb endal organiseerida.
2. Puudub tarkvarauuenduste automaatse levitamise süsteem.
3. MapReduce ei võimalda *in-memory* andmete töötlust, mis teeb selle märkimisväärselt aeglasemaks paljudest teistes paralleeltöötluse lahendustest.
4. Enamus tööst käib käsureal. Pole välja arendatud graafilist kasutajaliidest, kus programmeerimisoskusteta kasutaja saaks andmeid analüüsida.
5. Tasuta versioonide osas pole väljaarenenud paketeeritud paigalduskomplektide süsteem, mis teeb vajalike komponentide otsimise ja leidmise keeruliseks ning väga ajamahukaks.
6. Vaikimisi ei ole olemas võimalus autentimise ja autoriseerimise ühendamiseks kohaliku LDAP-iga.
7. Eestis puudub piisav spetsialistide võrgustik, kes oleks võimeline süsteemi paigaldama ja haldama.

## Arhitektuuri valimine

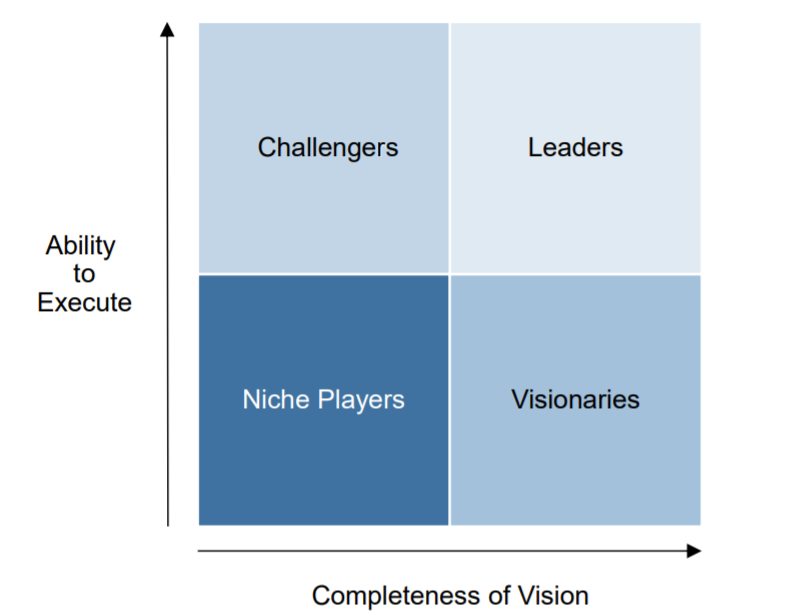
|  |  |  |  |
| --- | --- | --- | --- |
| Funktsio-naalsus | Variant 1 | Variant 2 | Variant 3 |
| Visualiseeri-mine ja andmete avaldamine | * Power BI Pro * Power BI Cloud * Azure Cloud | * Apache Zeppelin | * Appache Zeppelin * GeoSparkViz |
| Analüüs | * Power BI Pro * SQL Server Analysis Services * R * Python | * Apache Zeppelin * Apache MapReduce * R * Python | * Appache Zeppelin * Apache Spark * GeoSpark * R * Python |
| Andmete töötlemine | * SSIS * MS SQL Server 2019 Polybase * SSMS | * Apache Hadoop * Apache MapReduce | * Apache Spark |
| Andmete sisestamine ja haldamine | * Andmete sisestus- ja haldusmoodul (spetsiaalarendus) * MS Sharepoint * MS Teams * Office 365 | * Andmete sisestus- ja haldusmoodul (spetsiaalarendus) | * Andmete sisestus- ja haldusmoodul (spetsiaalarendus) |
| Andmete hoidmine | * MS SQL Server 2019 Polybase | * Apache Hadoop,HDFS * HBase | * Postgre * PostGIS |
| Rakenduste majutus ja haldamine | * Windows Server * Azure Cloud * Power BI cloud * (või privaatpilv) | * Linux * Apache Yarn * Google Cloud (või privaatpilv) | * Linux * Pgadmin * Google Cloud (või privaatpilv) |

**Tabel 12. Arhitektuurivariantide konfiguratsioonid**

Analüüsi käigus selgus, et andmemahud, mida Suurandmete süsteem peab käsitlema, on suurandmete definitsioonide mõistes väikesed. Andmete keerukuse kohta seda öelda ei saa. Samuti andmekvaliteedi probleemide kohta. Selle tõttu on esimeseks nõudeks arhitektuuri juures keerukuse käsitlemise võime ja keerukusega toime tulemise kiirus ehk tulemite saavutamise kiirus teenuste loomisel.

Arhitektuurivariantidest on kõige komplektsem variant sellises olukorras toime tulemiseks **variant 1.** Seda nii lokaalse paigalduse, kui ka pilvepaigalduse korral. Millel antud väide põhineb?

Uuringufirma Gartner Inc teeb iga-aastaseid uuringuid analüütika tarkvara pakkujate toodangu kohta. Hinnang tarkvarale paigutatakse nelja ruutu, mis on kujutatud järgneval joonisel:



**Joonis 12. Gartneri Magic Quadrant tarkvaratoodete hindamiseks, Allikas Gartner Inc.**

Tootjad jagatakse nelja klassi:

1. liidrid;
2. visionäärid;
3. nišimängijad;
4. väljakutsete esitajad.

Joonisel näidatud teljed näitavad horisontaalsuunas visiooni katvust erinevate tehnoloogia ja lahenduste osas ning vertikaalsel suunal rakenduse genereerimise võimet. Paremasse ülemisse kasti jäävad tootjad, kelle toodang võimaldab kasutajal luua tulemit kõige kiiremini ja maksimaalselt katva funktsionaalsuse abil.

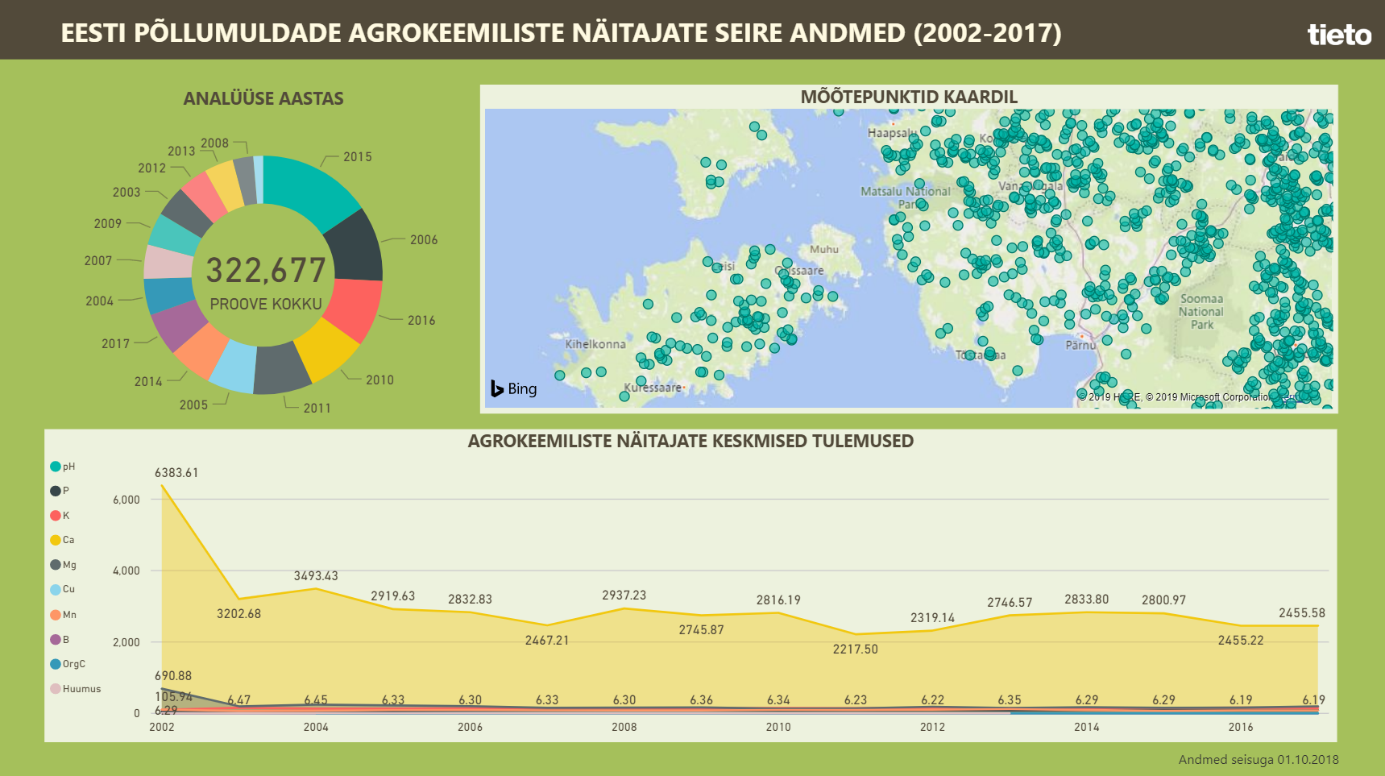


**Joonis 13. Hinnang andmeanalüütika lahenduste tootjatele, allikas Gartner Inc.**

Jooniselt on näha, et nii lahenduse visiooni, kui ka elluviimise ja käitamise osas on hetkel vaieldamatu liider Microsofti toodang (Power BI, SQL Server, SSIS, SSMS jt). Järgnevad Tabeleau ja Qlik. Microsofti kasuks räägib lisaks heale visioonile ja kõrgele efektiivsusele rakenduste tootmises ka madalad kulud, mis leidis tõendust käesoleva projekti käigus. Kulude kokkuhoiu oluline komponent oli väike halduse maht.

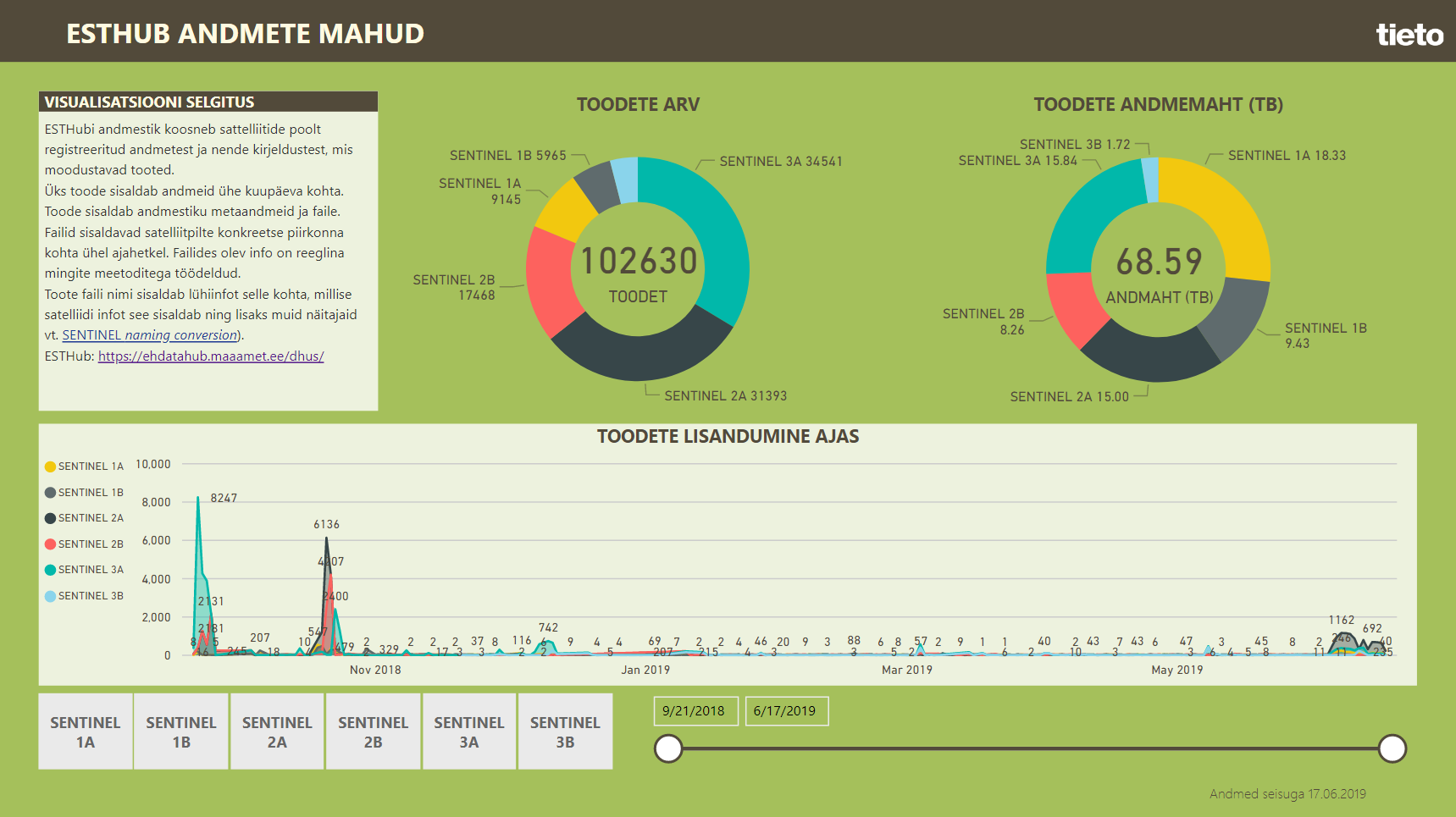
Esimest arhitektuurivarianti katsetati projekti analüüsistaadiumis, kus loodi kõiki andmekogusid kattev andmete kvaliteeti ja sisu visualiseeriv analüüsirakendus. Positiivseks oli selle juures kiiresti tekkivad tulemid, lõpptarbija rakenduse stabiilsus ja haldamise lihtsus ning haldamise väike töömaht. Projekt kestis 11 kuud ja selle aja jooksul süsteemi töös blokeerivaid katkestusi ei olnud. Kasutajate hulk oli analüüsi poolel 4 inimest ja visualisatsioone kasutas kümneid inimesi, kes projektis osalesid.

Variant 1 toodete puhul kardetakse tihti nende litsentside maksumust. Antud kasutajate hulga juures läks kogu süsteem paigaldatuna pilve otseste kuludena maksma ca 1200 eurot, mis on 11 kuu kohta väga väike summa arvestades, et kaasatud oli üle 40 andmekogu andmed, mitukümmend kasutajat ning nii SQL Server, Azure, Power BI ja SSMS tarkvarad. Süsteemi haldamiseks ei läinud tarvis eraldi administraatoritest koosnevat haldusüksust. Esimesed analüüsi tulemid, mis kujutasid endast interaktiivseid analüütikarakendusi, koostatuna ilma arendajate abita, valmisid juba projekti esimestel nädalatel.



**Joonis 14. Analüütikarakenduse näidis, mullastiku agrokeemiliste näitajate andmete kokkuvõte**

Rakendus kuvab mullaanalüüside andme agregeeritud kujul.

****

**Joonis 15. Analüütikarakenduse näidis, ESTHub andmete juurdekasvu monitoorimine**

Joonisel kujutatud rakendus võimaldab monitoorida ESTHub andmete juurdekasvu erinevatest dimensioonidest lähtuvalt.

Kõikide nimetatud kolme arhitektuurivariandi läbivaks jooneks on R-i ja Pythoni kasutamine analüüsialgorimide realiseerimisel. Seega on võimalik ühel platvormil loodud programme rakendada ka teistel. Kui orienteeruda ühele arhitektuurile, siis täiendavate asjaolude selgumisel on siiski võimalik arhitektuuris teha muudatusi kuni selleni, et minnakse üle teisele platvormile.

# Suurandmete süsteemi standardite kirjeldus

Suurandmete süsteemi standardid on määratletud selleks, et oleks olemas lähtepunkt, mille baasil tekiks tehniline koosvõime erinevate andmekogude vahel ning andmeteaduslik koosvõime seotud asutuste vahel. Ilma standardite kasutamiseta tekiks palju 1:1 kokkuleppeid ja koosvõime ning koostöö mitme osapoole vahel korraga on raske tekkima. Teadmuse ja andmete kasutamine teisest andmekogust on oluliselt raskendatud, kui ei ole järgitud kokkuleppeid, mis saavad alguse standarditest. Oluline on, et standardite kasutamise aluseks on olemas ka õiguslik kokkulepe.

## Füüsilisele andmemudelile esitatavad nõuded

1. Andmekogu andmemudel peab olema dokumenteeritud selliselt, et seda oleks võimalik masintöödelda. Laiemalt on levinud on UML formaat eap kujul. Seda toetab tööriist SPARX Enteprise Achitect. Maaeluministeeriumi uute infosüsteemide puhul on antud tööriist juba kasutusel ja tuleks rakendada ka mujal.
2. Andmemudel peab olema UML data model tüüpi ja seda peab olema võimalik XMI failina (<https://www.omg.org/spec/XMI>) avatud formaadis edastada teistesse metaandmete ja andmemudelite haldamise lahendustesse.
3. Juhul, kui ajakohast füüsilist andmemudelit ei ole olemas, tuleb see luua reaalandmebaasist pöördprojekteerimise teel nii, et mudel kirjeldaks reaalset füüsilist andmemudelit, mitte kellegi subjektiivset nn loogilist mudelit.
4. Tabelid peavad olema kirjeldatud järgmiselt:
   * Andmetabeli nimi peab olema eesti või inglise keeles ja väljendama võimalikult lähedaselt tabeli sisu.
   * Andmetabelil peab olema kommentaar (Notes), kus kirjeldatakse pikemalt, kui nimes, mida andmetabelisse salvestatakse.
5. Tabelite veerud peavad olema kirjeldatud järgmiselt:
   * Tabeli veerul peab olema veeru andmesisu väljendav eesti või inglise keelne nimi.
   * Tabeli veerul peab olema kommentaar (Notes), mis iseloomustab veeru andmesisu. Juhul, kui veergu salvestatakse klassifikaatori elemendi koode, siis peab olema kirjeldatud klassifikaatori kood, mille elemente veerg kasutab.
   * Veeru andmetüüp peab olema määratud andmetüüpide hulgast, mida võimaldab kasutada andmebaasi mootor, milles andmebaasi hoitakse. See ühtlasi tähendab ka seda, et igal tabelil peab olema määratletud andmebaasimootori nimetus.
   * Kohustuslikud veerud peavad olema tähistatud.
6. Tabelite vahelised seosed peavad olema defineeritud nii, et oleks aru saada milliselt väljalt millisele viidatakse ja millist tüüpi seos on (0..n, 0..1, 1..1, n..n).
7. Füüsiliselt andmemudel peab olema selline, et sellest oleks võimalik genereerida andmebaasi loomise DDL-skript. Igal ajahetkel peab olema võimalik kogu andmebaas uuesti luua ilma süsteemi arenduse käigus loodud inkrementaalsete muudatuste rakendamise skripte kasutamata.

## Mõõdikute kirjeldusele esitatavad nõuded

Mõõdikud väljenduvad andmebaasides andmeelementidena, mis iseloomustavad reaalse maailma nähtusi kvantitatiivsetest ja kvalitatiivsetest aspektidest. Mõõdik on üks Suurandmete süsteemi poolt antava väärtuse alustalasid, millest kõik alguse saab. Defineerimata sisuga andmete väärtus on täppispõllumajanduses küsitav. Hoolimata sellest, et mõned andmed on kirjeldamata, jõutakse iga andmestikuga alati selleni, et nende sisu tehakse kindlaks ja sisu kirjeldus võetakse aluseks erinevatele järgnevatele töödele ning teenustele. Andmete puhul peab olema määratletud nende sisu ja selle juures ka andmete kogumise metoodika, et oleks saavutatav andmete võrreldavus. Andmete sisu tähendus oleneb nii mõõtmisel kasutatud seadmest või andmekogumise viisist, kui ka muudest dimensioonidest, mis mõõtmise hetkel kasutusel olid. Teadmata neid üksikasju võib väetussoovitus, põllumajandusmasina töö juhtandmestik või huumusbilanss olla väära sisuga. Suurandmete süsteemi peamine ülesanne on kasutades minevikus registreeritud andmed prognoosida tulevikku. Soovitud tuleviku ja eesmärkide saavutamine maksimaalselt optimaalsete ressurssidega on kõikvõimalike tööplaanide ja prognooside koostamise mõte.

Mõõdikutega käivad käsikäes dimensioonid, mis annavad mõõdikutele koha ajas ja ruumis ning täiendavates kategooriates, mis väljendatakse klassifikaatorite ja koodilistide kujul. Näiteks põldkatse mõõdikud on saak ning väetise kogus ja dimensioonideks koht ja aeg, kus katse toimus. Dimensioonina käsitletakse ka objekte ning subjekte, mis katsega seotud olid.

Mõõdikute kirjelduses kasutame järgmisi andmeelemente:

| Element | Kirjeldus |
| --- | --- |
| Kood | Mõõdiku unikaalne kood üle kõikide süsteemis kirjeldatud mõõdikute. Kuna paljud mõõdikud on sarnaste nimedega, siis on mõistlik mõõdikud kodeerida. |
| Nimi | Mõõdiku nimi, mis iseloomustab mõõdiku sisu. Mõõdiku nimi võib anda viiteid ka dimensioonidele, millega mõõdik seotud on. Näiteks „Nisu saak“, „Õhutemperatuur“, „Mulla niiskus 30cm sügavusel“. |
| Kirjeldus | Mõõdiku kirjeldus, mis iseloomustab mõõdiku sisu täpsemalt. |
| Mõõtühik | Mõõdiku puhul kasutatav mõõtühik. Näit. kg, t, s. |
| Andmetüüp | Mõõdiku andmetüüp. Võimalikud väärtused on „tekst“, „täisarv“, „reaalarv“, „klassifikaator“. |
| Formaat | Mõõdiku formaat. Näit. „xls“, „DDMMYYYY“. Siin näidatakse ära ka see, mitme komakohaga on reaalarv, kui see on andmestikus täpselt määratletud. |
| Grupp | Mõõdiku grupi kood. Kuna mõõdikuid on tuhandeid, siis tuleb neid kuidagi grupeerida, et üldpilt ja õige mõõdiku leidmise võimalused oleksid suuremad. |
| Haldaja | Kes mõõdiku kirjeldust haldab. Siin ei peeta silma mitte seda, kes mõõdiku andmed ehk näitajad tekitab vaid asutust või isikut, kes vastutad mõõdiku kirjelduse haldamise eest. See tähendab ühtlasi ka seda, selle määratlemist, milline on mõõdiku andmete (näitajate) kogumise metoodika. |
| Näidisandmed | Näidisandmed, mis iseloomustavad täiendavalt mõõdiku sisu. |
| Andmete allikad | Mõõdiku seos andmeallikatega. Siin võib olla näidatud mitu andmeallikat (andmeallika koodi). |
| Kas mõõdik on algandmestik või tuletis | Iga mõõdik peab olema määratletud kas algandmetena või algandmetest arvutatud mõõdikuna. Võimalikud väärtused siin on „algandmed“ või „tuletis“. Kui mõõdiku puhul on tegemist agregeeritud andmetega, siis seda nimetatakse alati tuletiseks. Tuletis võib tekkida arvutamisel algandmetest, millele on rakendatud mitte liitmistehe vaid ka mõni keerukam valem. |

**Tabel 13. Mõõdiku kirjelduse struktuur**

## Dimensioonide kirjeldusele esitatavad nõuded

Dimensioonidel andmestik on tavaliselt selline, mis end ise selgitab. Selle tõttu on oluline dimensiooni enda kvaliteet. Siiski, et oleks teada, millised dimensioonid kasutusel on, tuleks neid mõne sõnaga iseloomustada ja süstematiseerida.

Klassifikaatorid on oma olemuselt dimensioonid, kuid kõik dimensioonid ei ole klassifikaatorid. Näiteks aja dimensioon.

|  |  |
| --- | --- |
| Element | Kirjeldus |
| Kood | Dimensiooni unikaalne kood üle kõikide süsteemis kirjeldatud dimensioonide. |
| Nimi | Dimensiooni nimi, mis iseloomustab dimensiooni sisu. Näit. „Mõõtmisviis“, „Katseaasta“, „Aeg“. |
| Kirjeldus | Dimensiooni lühikirjeldus, mis iseloomustab dimensiooni sisu täpsemalt. |
| Grupp | Dimensiooni grupi kood. Kuna dimensioone on tuhandeid, siis tuleb neid kuidagi grupeerida, et üldpilt ja õige mõõdiku leidmise võimalused oleksid suuremad. |
| Haldaja | Kes dimensiooni kirjeldust haldab. |
| Klassifikaatori kood | Juhul, kui dimensiooni näol on tegemist klassifikaatoriga, siis tuleks siia näidata klassifikaatori kood. |

**Tabel 14. Dimensiooni kirjelduse struktuur**

Tabelis on kirjeldatud dimensiooni kirjeldamisel kasutatavad andmeelemendid.

## Klassifikaatoritele esitatava nõuded

Klassifikaatorid peavad vastama järgmisele struktuurile:



**Joonis 16. Klassifikaatorite andmete struktuur**

Klassifikaatorite haldamise lahenduse juures tuleb arvestada järgmiste nõuetega:

1. Klassifikaatorite haldamine on oma iseloomult andmete masstöötlus. See tähendab, et klassifikaatorid ja nende elemendid saabuvad süsteemi ning ka väljastatakse süsteemist massandmetena kas CSV või XLSX kujul. Suurandmete süsteemis peab olemas olema vastav funktsionaalsus.
2. Klassifikaatoreid peab olema võimalik süsteemis võrrelda, et oleks võimalik leida korduvaid klassifikaatoreid. Võrdluse tulemuse peab süsteem looma klassifikaatorite elementide vahele seoseid st võrdluse tulemuse salvestama.
3. Klassifikaatoreid peab saama ühendada ja lahutada erinevateks klassifikaatoriteks.
4. Suurandmete süsteemil peavad olemas olema klassifikaatorite andmete veebiteenused, mille kaudu saavad tarbijad klassifikaatorite andmeid pärida ning vastavalt õigustele ka uuendada.

### Tabel: KLASSIFIKAATOR

Andmekogus kasutatav klassifikaator. Võib olla ka koodiloend, st ei pruugi katta kogu vaadeldavat andmeobjektide kogumit.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Primaar-  võti | Atribuudi nimi | Vaike-väärtus | Andme-  tüüp | Kohustus-  likkus | Kirjeldus |
| True | KOOD |  | varchar(100) | True | Klassifikaatori kood või lühend. |
| False | NIMI |  | varchar(50) | True | Klassifikaatori nimi. |
| False | KIRJELDUS |  | varchar(4000) | False | Klassifikaatori kirjeldus. |
| False | HALDAJA\_KOOD |  | varchar(50) | False | Klassifikaatorit haldava asutuse registreerimisnumber.. |
| False | ON\_KLASSIFIKAATOR |  | int | False | Juhul, kui 1, siis on tegemist klassifikaatoriga. Kui on 0, siis ei ole tegemist klassifikaatoriga vaid koodilistiga, mis ei pruugi katta kõiki vaadeldavaid objekte. |
| False | VALDKOND\_KOOD |  | varchar(100) | False | Millise valdkonna kohta klassifikaator käib - viide valdkonna klassifikaatori elemendile. |
| False | LISAATRIBUUDID |  | xml | False | Klassifikaatori elemendi lisaatribuudid kujul:  <COLOR>RED</COLOR>  <SIZE>BIG</SIZE>  <SUBJECT>DATA</SUBJECT> |

**Tabel 15. Klassifikaatori andmete struktuur**

### Tabel: KL\_ELEMENT

Klassifikaatori või loendi element.

| Primaar-  võti | Atribuudi nimi | Vaike-väärtus | Andme-  tüüp | Kohustus-  likkus | Kirjeldus |
| --- | --- | --- | --- | --- | --- |
| True | KOOD |  | varchar(100) | True | Klassifikaatori elemendi kood. |
| False | NIMI |  | varchar(200) | True | Klassifikaatori elemendi nimetus. |
| False | TASE |  | int | False | Elemendi taseme number, kui tegemist on hierarhilise klassifikaatoriga. |
| False | YLEM\_KOOD |  | varchar(100) | False | Ülemelemendi kood, kui on tegemist hierarhilise klassifikaatori hierarhias alumistel tasemetel paikneva elemendiga. |
| False | KIRJELDUS |  | varchar(4000) | False | Klassifikaatori elemendi kirjeldus, mis annab selgituse, millistel juhtudel vaadeldav objekt antud elemendi alla kuulub. |
| False | KEHTIVUSE\_ALGUS |  | date | False | Elemendi kehtivuse algus. |
| False | KEHTIVUSE\_LOPP |  | date | False |  |
| False | MOOTYHIK\_KOOD |  | varchar(100) | False | Viide mõõtühikute klassifikaatori elemendile. |
| False | KL\_VERSIOON\_KOOD |  | varchar(100) | False |  |
| False | LISAATRIBUUDID |  | xml | False | Klassifikaatori elemendi lisaatribuudid xml kujul, näiteks:  '<Omadused>  <Värv>kollane<Värv>  <Pikkus>200</Pikkus>  </Omadused>' |
| False | KLASSIFIKAATOR\_KOOD |  | varchar(100) | True | Viide klassifikaatorile, millesse element kuulub. |

**Tabel 16. Klassifikaatori elemendi andmete struktuur**

### Tabel: KL\_VERSIOON

Klassifikaatori versioon.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Primaar-  võti | Atribuudi nimi | Vaike-väärtus | Andme-  tüüp | Kohustus-  likkus | Kirjeldus |
| True | KOOD |  | varchar(100) | True | Klassifikaatori versiooni kood. |
| False | NIMI |  | varchar(200) | True | Klassifikaatori versiooni nimi. |
| False | KIRJELDUS |  | varchar(4000) | False | Klassifikaatori versiooni kirjeldus. |
| False | KEHTIV |  | int | True | Kas klassifikaatori versioon on kehtiv ('1') või kehtetu ('0'). |
| False | KEHTIVUSE\_ALGUS |  | date | True |  |
| False | KEHTIVUSE\_LOPP |  | date | False |  |
| False | TYYP |  | varchar(50) | True | Kas klassifikaatori versioon on 'aegpidev' või 'versioneeritav'. NB! Klassifikaatoril võib korraga olemas olla nii aegpidev versioon , kui ka muu versioon. |
| False | KLASSIFIKAATOR\_KOOD |  | varchar(100) | True | Viide klassifikaatorile. |
| False | YLEMINEKUTABEL |  | xml | False | Ülemineku tabel või tabelid xml kujul. Üleminekutabelis on näidatud, kuidas on võimalik teisendada klassifikaatori elemendid ühest versioonist teise. |

**Tabel 17. Klassifikaatori elemendi andmete struktuur**

### Tabel: SEOS

Klassifikaatori elementide vaheline seos. Seose tüüpe võib olla erinevaid ja neid peab saama ka lisada.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Primaar-  võti | Atribuudi nimi | Vaike-väärtus | Andme-  tüüp | Kohustus-  likkus | Kirjeldus |
| False | KL\_ELEMENT\_KOOD\_1 |  | varchar(100) | True | Klassifikaatori elemendi esimese seose osapoole kood. |
| False | KL\_ELEMENT\_KOOD\_2 |  | varchar(100) | True | Klassifikaatori elemendi seose teise osapoole kood. |
| False | SEOS\_TYYP |  | varchar(100) | True | Seose tüübi tähis. |
| False | KOMMENTAAR |  | varchar(4000) | False | Vajadusel kirjeldatakse siia seose olemus. |

**Tabel 18. Klassifikaatori elemendi andmete struktuur**

## Riiklikud nõuded metaandmete kirjeldamise osas

Riiklikul tasemel on andmehalduse korraldaja Statistikaamet, kelle poolt on koostatud andmehalduse juhtimise tegevuskava (vt [[2]](https://www.stat.ee/dokumendid/1413034)). Statistikaameti poolt on väljatöötamisel andmekirjelduse standard. Ka põllumajanduse Suurandmete süsteemi metaandmed peavad olema kooskõlas selle standardiga. Vorming metaandmete käsitlemiseks on Statistikaameti poolt planeeritud DDI standard (vt. [[3](https://www.ddialliance.org/)]).

## Teenuste kirjeldustele esitatavad nõuded

Suurandmete süsteemi avaliku teenuste loomine eeldab teenuse kirjelduse olemasolu. Kõige olulisem iga teenuse osas on väljundmõõdikute kirjeldus ehk see, millist infot antud teenus toodab. Samasugune loetelu tuleb määratleda kas sisendmõõdikute osas. Sisendid määravad paljuski teenuse loomise maksumuse. Teenust ei ole võimalik luua, kui teenuse sisendandmeid ei ole kusagilt võtta. See kehtib eriti just massandmeid tarbivate teenuste puhul. Lihtne kalkulaator, kus kasutaja sisestab teenuse kasutamisel selle sisendandmed, ei ole Suurandmete süsteemi tööobjekt. Suurandmete süsteemi teenuseks muutub teenus siis, kui on tarvis sisendis kasutada ühte või mitut andmeallikat, kust on võimalik massandmeid kätte saada automaatselt üle veebiteenuse või siis regulaarselt faili või andmebaasiühenduse kaudu.

Et Suurandmete süsteemi teenustest oleks selge ülevaade, tuleb süsteemi kirjeldada teenuste sisu, mis koosneks järgnevast:

| Element | Kirjeldus |
| --- | --- |
| Kood | Teenuse unikaalne kood üle kõikide süsteemis kirjeldatud teenuste. |
| Nimi | Teenuse nimetus. Näiteks „Huumusbilansi kalkulaator“. |
| Liik | Teenuse liik. Valikuvariandid on:   1. „Kasutajateenus“ – teenus, mille kaudu saab . 2. „Kasutajaliides“ - nähtav ja kasutatav ligipääs kasutajale, st teenus mille kaudu kasutaja saab andmeid vaadata ja kasutada. 3. „Veebiteenus“ – *web service* ehk infosüsteem-infosüsteem taseme masinliides, mida kasutades saab infosüsteem ligipääsu andmeallika andmetele. 4. „Andmebaas“ – andmebaas või avaandmete fail, andmete kogum, millel puudub spetsiaalne kasutajale suunatud juurdepääsuteenus. |
| Kirjeldus | Teenuse lühikirjeldus, mis iseloomustab teenuse sisu täpsemalt sh kirjeldab ära, miks teenus luuakse, mis on saavutatav kasu, samuti kes on kasutajad. |
| Sisendid ja väljundid | Sisendite ja väljundite kirjeldus. Siin peab olema toodud sisendandmete ja väljundandmete kirjeldus mõõdikute ning dimensioonide kirjeldusena. |
| Grupp | Teenuste grupi kood. Teenused on mõistlik grupeerida temaatilistesse gruppidesse. Nii on lihtsam teenuste kirjeldusi süstematiseerida ja leida dubleerivaid kirjeldusi. |
| Omanik | Teenuse omanik on asutus või isik, kes tegeleb teenuse metoodilise sisu püstitamise ja arendamisega. |
| Teenuse haldaja | Teenuse tehnilise toimimise eest vastutav asutus. |
| Püstitaja | Isik või asutus, kes teenuse loomise idee algatas. |
| URL | Teenuse link. |

**Tabel 19. Teenuse kirjelduse struktuur**

Teenused kirjeldatakse Suurandmete süsteemis sarnaselt, et oleks võimalik moodustada teenuste kataloog ning selle efektiivne otsingusüsteem.

## Täiendavad soovitused kirjelduste koostamiseks

Kirjelduste puhul tuleb täiendavalt jälgida järgmisi punkte:

1. Kirjeldused ei tohi üksteist dubleerida.
2. Kirjeldusi tuleb hallata ühtses andmebaasis.
3. Teenuste kodeerimisel ei ole soovitav kasutada ainult numbrilisi koode vaid võimaluse korral lisada koodile ka objekti iseloomustav lühend, et kood oleks loetavam ja tekiks vähem eksimusi koodide kasutamisel.
4. Kõikidel kirjeldusel peab olemas olema OData veebiteenus, mis võimaldab kirjelduste infot kasutada loodavates teenustes ja andmeanalüüsides.

## Veebiteenuste standardid

Järgnevas on kirjeldatud Suurandmete süsteemis kasutatavate veebiteenuste liigid.

| ****Kood**** | ****Nimi**** | ****Kirjeldus**** | ****Viited**** |
| --- | --- | --- | --- |
| WMS | Web Map Service | Geoinfo vahetamine pildi kujul. Teenuse liik on mõeldud kaartide pildina alla laadimiseks. | <http://www.opengeospatial.org/standards/wms> |
| WFS | Web Feature Service | Geoinfo vahetamine tärkkujul, kus edastatakse info *feature* ja *feature protperty* tasemel. Teenuseliik on kasutusel geoinfo edastamiseks detailsel kujul, kus tarbija saab ise otsustada, kuidas infot lõpuks kuvada. | <http://www.opengeospatial.org/standards/wfs> |
| WCS | Web Coverage Service | Geoinfo vahetamine multidimensionaalsel kujul, kus päritavad dimensioonid on võimalik pärijal ette anda. Erineb teistest geoinfo teenustest selle poolest, et võimaldab pärida infot osade kaupa. | <http://www.opengeospatial.org/standards/wcs> |
| Point Cloud teenus | Punktipilve teenus | Punktipilve teenus on mõeldud ruumiinfo vahetamiseks punktipilvena. Iga punkti kohta antakse edasi X, Y, Z koordinaadid ja seda punkti iseloomustavad muud näitajad, mida võib olla n tükki. Teenuseliik on mõeldud kasutamiseks Lidari andmete ja teiste andmete puhul, kus mõõtmise täpsus on punkt, mitte ruumiline areaal. Selline struktuur on kasutusel näiteks maapinna kõrgusmudelis. Sama struktuuri tuleks rakendada ka mulla koostise jt täppispõllumajandust toetatavatel andmestikel. | <https://en.wikipedia.org/wiki/Point_cloud> |
| RWS | RESTful veebiteenus | Veebiteenuste liik, kus teenusega toetatavad operatsioonid on standardiseeritud ja info liigub teksti kujul. RESTful teenused on kasutatavad eriti avaandmete edastamiseks, kus ei ole oluline kasutaja sessioon. | <https://en.wikipedia.org/wiki/Representational_state_transfer> |
| OData | OData RESTful veebiteenus | Üks laiemalt toetatud REST teenuste standardeid avaandmete jagamiseks on Open Data veebiteenused. See on välja töötatud OASIS-e poolt ning kinnitatud ISO/IEC JTC 1 poolt. OData adapterid on olemas ka paljudel andmeanalüüsitarkvaradel nagu Power BI ja Tableau, mis loob OData standardi järgi loodud teenustele laiapõhjalise kasutajaskonna. Samuti ei ole keeruline OData teenust kasutusele võtta ükskõik millises spetsiaaltarkvaras.  OData teenuse andmestruktuuri planeerides tuleks arvestada järgmiste asjaoludega:   1. Teenuse andmestruktuur ei tohi olla liiga sügav ega ka rekursiivne, et teenust oleks võimalik ilma kasutada universaalsete adapteritega laiatarbe analüüsitarkvarades. Andmeid peab olema võimalik teisendada kahemõõtmeliseks tabeliks ehk denormaliseerida ilma kirjete arvu eksponentsiaalselt suurendamata. 2. Juhul, kui denormaliseerimine muudab kirjete arvu liiga suureks (>5 000 000), tuleb teenus jagada mitmeks tükiks nii, et kõiki andmeobjekte ei paigutata ühte struktuuri vaid tekib osaliselt denormaliseeritud struktuur, kus eksisteerib seoseid kahe erineva teenuse andmete vahel. 3. Teenuse kasutatavust tuleks testida laiatarbe analüüsitarkvaradega, et tuvastada kriitiline andmestruktuuri keerukus, mille denormaliseerimisele tarkvara vastu peab. | <https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=odata>,  <https://www.oasis-open.org/news/pr/iso-iec-jtc-1-approves-oasis-odata-standard-for-open-data-exchange>,  <https://www.odata.org/> |
| X-tee teenus | X-tee teenus | Eesis kasutatav SOAP protokollil põhinev teenusliik autentimist ja autoriseerimist ning vastavat infrastruktuuri nõudev teenuse liik. Andmestruktuurid defineeritakse X-tee teenuste puhul ette ning operatsioonide hulk ei ole piiratud. X-tee teenused on põhiliselt kasutusel konfidentsiaalse info liigutamiseks. | <https://www.ria.ee/et/riigi-infosusteem/andmevahetuskiht-x-tee.html> |

**Tabel 20. Veebiteenuste standardid**

Tuleb arvestada, et paljudel andmekogudel on teenused juba loodud ning sel juhul on majanduslikult otstarbekam luua vastavad adapterid, mitte hakata teenuseid juurde looma. See kehtib eriti valmistarkvarade puhul, mis on kasutusel näiteks agrometeoroloogias ja ka põllumajandusmasinates.

## Andmetüüpide kasutamine

Süsteemi kirjelduste tasemel kasutatavad (semantilised) andmetüübid on kirjeldatud järgnevas tabelis:

|  |  |
| --- | --- |
| Andmetüüp | Selgitus |
| Täisarv | Ilma komadeta numbriline arv. |
| Reaalarv | Komakohtadega arv. |
| Tekst | Suvalise suurusega tekst. Tekst võib sisaldada ka kuupäeva või aega. |
| Ruumikuju | Nähtuse ruumikuju iseloomustav andmemassiiv. Laiemalt tuntud, kui andmetüüp *geometry*. Ruumikuju tüüp on täpsemalt kirjeldatud [[6](https://www.postgresql.org/docs/9.4/datatype-geometric.html)]. |

**Tabel 21. Teenuse kirjelduse struktuur**

Füüsiliste andmemudelite ja andmebaaside osas kasutatakse konkreetse andmebaasimootori põhist andmetüübistikku.

## Andmekvaliteedi standard

Suurandmete süsteemi teenuste kvaliteedi tagamiseks peab iga liidestatav teenus olema läbinud andmekvaliteedi analüüsi.

Andmekvaliteedi analüüs tuleb läbi viia reaalandmete baasil ning hinnata tuleb järgmises tabelis toodud aspekte.

| Kood | Hinnatav aspekt | Selgitus |
| --- | --- | --- |
| AKK-1 | RIHA andmekoosseisu vastavus reaalandmebaasile | Kas andmekoosseisu kirjeldus, mis paikneb RIHA-s, vastab reaalsele andmebaasile? Kas kirjeldus vastab andmekogu põhimäärusele? Ebalegaalselt kogutud andmetele ei saa rajada Suurandmete süsteemi teenuseid. Juhul, kui teenus luuakse tuleb samaaegselt korrastada ka andmete õiguslik taust. |
| AKK-2 | Andmete kirjelduse olemasolu | Leitakse vastus küsimusele, milline on reaalne andmekogu andmestik. Kas andmemudel on dokumenteeritud. |
| AKK-3 | Andmete kogumise metoodika olemasolu ja järjepidevus | Millise metoodika järgi andmeid kogutakse sh kas on olemas metoodiline järjepidevus? |
| AKK-4 | Andmete vormi järjepidevus | Kas andmed on esitatud vormiliselt järjepidevalt? Kas andmetest on võimalik moodustada aegridu. |
| AKK-5 | Andmete usaldusväärsus | Kas andmed on usaldusväärsed? Kas andmekvaliteedi tagamiseks on andmekogus võetud tarvitusele meetmeid? |
| AKK-6 | Andmete sisutihedus | Kas andmed on esitatud sisutihedalt st kas andmestruktuur sisaldab tühje veerge ja tabeleid ning kas andmebaasis olevad andmed on mahuliselt optimaalselt esitatud? |
| AKK-7 | Andmete ülekande võimalikkus | Kas andmed on kadudeta ülekantavad teise keskkonda? |
| AKK-8 | Andmete masintöödeldavus | Kas andmed on masintöödeldavad? |
| AKK-9 | Andmete jälgitavus | Kas andmed on andmekogus jälgitavad ehk kas nende abil on võimalik aru saada protsessidest, mida andmed kajastavad nii ajalises, kui sisulisest aspektist? |
| AKK-10 | Andmete terviklus | Kas andmekogu andmete terviklus on tagatud andmebaasi tasemel või muude meetoditega? |
| AKK-11 | Andmete kättesaadavus | Kas andmed on kättesaadavad? Kas andmete edastamise teenused on piisava käideldavuse klassiga? |
| AKK-12 | Andmete õigsus | Analüüsitakse vastavalt võimalustele ja info kättesaadavusele sh teistest andmekogudest, kas andmed on täpsed ehk kas need kajastavad reaalse maailma objekte, subjekte ja sündmusi õigesti? |
| AKK-13 | Andmete võrreldavus teiste allikate andmetega | Kas analüüsitava allika andmed vastavad standarditele, mis on Eestis kehtestatud sh kas isikukoode ja aadressandmeid kasutatakse samas formaadis, mis teised andmekogud? |
| AKK-14 | Klassifikaatorite ja koodilistide kasutus | Millised klassifikaatorid on andmetes kasutusel, sh:   1. Kas klassifikaatorite haldamise lahendus on olemas? 2. Kas riikide klassifikaator on kasutusel? 3. Kas ADS on kasutusel ja protsentuaalselt mil määral? 4. Kas andmestikus kasutusel olevad klassifikaatorite elemendid on olemas ka klassifikaatoris? 5. Kas klassifikaatoreid kasutatakse järjepidevalt? 6. Kas kasutatavad klassifikaatorid on aegpidevad või versioneeritud? |
| AKK-15 | Registrinumbrite unikaalsus | Kas andmekogus kasutatakse isikukoode, registrinumbreid ja kas need on korrektsed ning unikaalsed? |
| AKK-16 | Objekti või subjektiga seotud sündmuste õige järjestus | Kas andmeobjektiga seotud sündmused on loogilises ja ajalises järjestuses, näiteks kas looma sünni ja surma kuupäevad on ajaliselt õiges järjekorras? |
| AKK-17 | Andmete täielikkus | Kuidas on andmebaasi tasemel tagatud kohustuslike andmete olemasolu? |
| AKK-18 | Koguseliste andmete õigsus koguse mõõtmisel | Kas koguselised näitajad on positiivse väärtusega? |
| AKK-19 | Duplikaatide olemasolu | Kas andmetes esineb duplikaate ehk kas registris esineb mitu samaväärset kirjet ühe ja sama reaalse maailma objekti või subjekti kohta? |
| AKK-20 | Kirjete staatuste olemasolu | Kas kirjete staatuseid ja olekuid kasutatakse korrektselt, kas staatuse või oleku andmed on alati kirjes olemas? |

**Tabel 22. Andmekogude andmete kvaliteedi hindamise kriteeriumid**

Nimetud aspektid põhinevad osaliselt RIA koostatud uuringul [[4](https://www.ria.ee/sites/default/files/content-editors/publikatsioonid/andmekvaliteedi_uuringu_lopparuanne.pdf)].

## Andmekogu kvaliteedi hindamine

Enne, kui andmekogu teenused liidestatakse suurandmete süsteemiga peab toimuma andmekogu andmekvaliteedi hindamine. Hindamise tulemusest andmekogu kas liidestatakse või ei liidestata Suurandmete süsteemiga. Juhul, kui kvaliteet on vastavuses andmete kasutamise eesmärkidega, tuleb sisse seada andmekvaliteedi monitoorimise süsteem. Esialgsel kvaliteedi hindamisel loodud kontrollid on monitooringu süsteemi aluseks, mis tähendab, et sama hindamist hakatakse tegema perioodiliselt.

Kvaliteedi monitoorimise süsteemile esitatavad põhinõuded on:

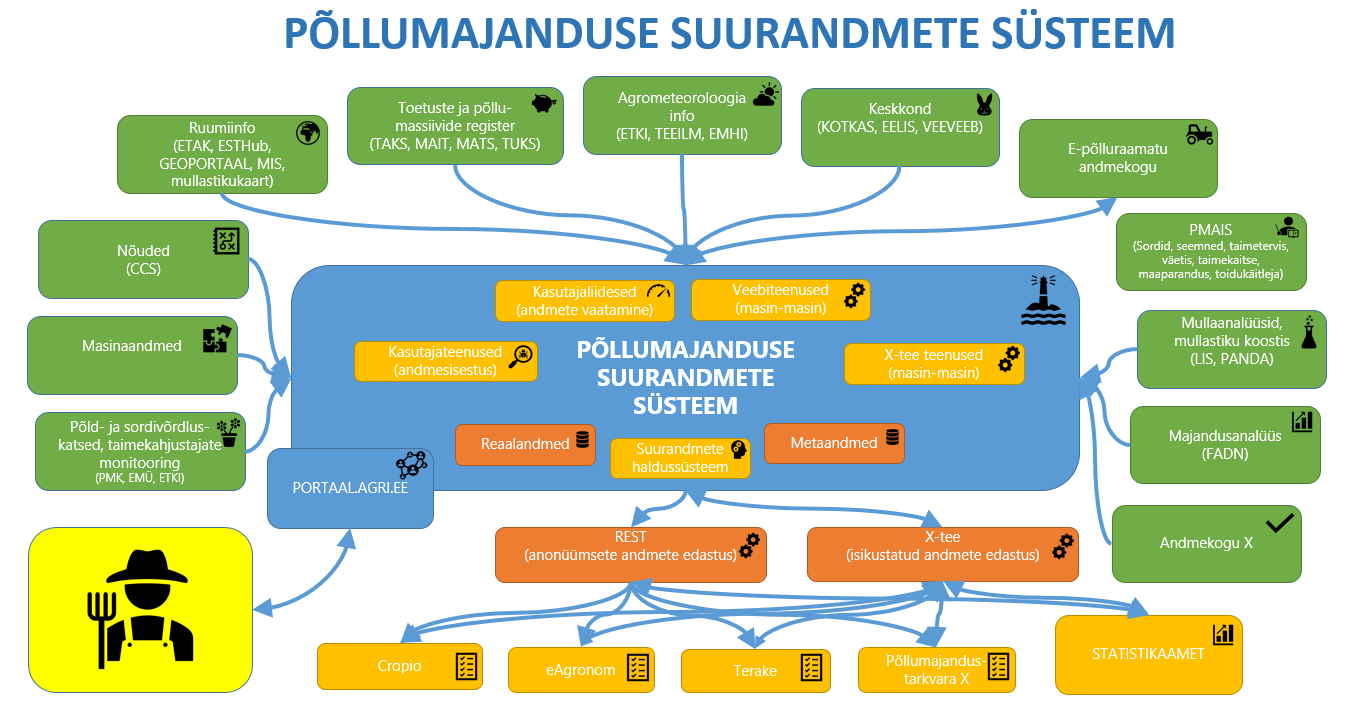
1. Süsteem peab toetama kontrolli kirjeldamisel R skripte.
2. Süsteem peab võimaldama koostada allika andmete peale andmemudeleid, millele kontrolle rakendada arvestades ka tabelite vaheliste seostega.
3. Süsteem peab võimaldama koostada ja rakendada kontrolle ühe tabeli ja kirje piires ning ka mitme tabeli piires, kasutades mitme tabeli veerge üheaegselt.
4. Kontrollimise tulemusi peab saama salvestada süsteemi, lisaks samas süsteemis ja veebis publitseerida kasutades levinud analüütilisi visualisatsioonide elemente (diagrammid, graafikud, tabelid).
5. Süsteem peab võimaldama kuvada nii algandmeid, kui neile rakendatud kontrollide tulemusi.

# Suurandmete süsteemi teenused

## Teenuste valik

Konsortsiumi liikmete poolt kaardistati analüüsi käigus üle 60 võimaliku teenuse. Tekkis olukord, kus klassifitseerimine ei võimaldanud sisuliselt tuvastada sotsiaalmajanduslikku kasutamise võimalusi. Täpsema vaatlemise käigus leiti, et teenused on otstarbekas täiendavalt grupeerida. Määratleti temaatilised grupid, hinnati esmaselt teenuse loomiseks vajalikud sisendid ja kirjeldati nn teenuste ökosüsteem. Sealhulgas eristusid teenused, millel ei ole kehtivat praktikat, et koostada konkreetne tuluhinnang, kuid mis on konsortsiumi poolt hinnatud kui eluliselt oluline komponent suurandmete süsteemi kui terviku toimimiseks.

Alljärgnev joonis kirjeldab põllumajanduse suurandmete tervikkontektsti, kuhu loodavad teenused suhestuvad ja väljendab seoseid, mida on vajalik arvesse võtta teenuste liigitamisel.



**Joonis 17. Põllumajanduse suurandmete süsteem**

Lähtuvalt eelnevast teenused grupeeriti alljärgnevalt:

* **Baaskompoenentide ökosüsteem** kirjeldab Suurandmete süsteemi jaoks olulised komponendid, millele konsortsium tervikuna andis kõige kõrgemad hinnangud. Samuti käsitleb baaskomponentide ökosüsteem kõige kõrgema hinnangu saanud tehnilisi teenused kui eeldusi, et uute teenuste turule teke oleks üldse võimalik. Majandusanalüüsi valiti eelkõige baaskomponentide alla kuuluvad teenused, kuid mis ei ole tehnilised teenused.
* **Tootmisteenused** kirjeldab olulised teenused, mis kuuluvad baaskomponentide alla, kuid on  sisulised teenused põllumajandustootmisega seonduvalt ning milledel on eristatav majanduslik tulu- ja kulu komponent. Majandusanalüüs viidi läbi just selle grupi teenustele.
* **Kaugseireteenused** kirjeldab gruppi kaugseire teenuseid, millest otseses praktilises kasutuses on vaid üks teenus, kuid millel on teaduspoole taustinfo ja laiapõhjaline potentsiaal teenuste valiku laiendamiseks. Kahjuks on keerukas praktika puudumise tõttu otsest väärtust turuosaliste poolt hinnata.
* **Masinandmete teenused**kirjeldab teenust, millel on seoseid nii paljude määratletud teenuste sisendite automatiseerimisel, mis on seni EL tasemel lahenduseta, mis on õiguslike regulatsioonide mõttes keerukas, kuid mille teostumisel tekib suur ja laiapõhjaline grupp potentsiaalseid kasutegureid.

## Teenuste majandusanalüüs

Käesoleva projekti “Teadmussiirde pikaajaline programm põllumajanduse suurandmete tegevusvaldkonnas” üks eesmärkidest on analüüsida, milline kasu tekib põllumajandusvaldkonna suurandmete süsteemi loomisest, millised on süsteemi loomise kulud ning kui otstarbekas on majanduslikus vaates süsteemi arendada. Majanduslik vaade tugineb tõhususe ja mõjususe kasutegurite hindamisele.

Suurandmete süsteemi majandusliku otstarbekuse tuvastamiseks viidi läbi majandusanalüüs, mis käsitles järgmisi aspekte:

* Millised on Suurandmete süsteemi teenused ja nende omavahelised sõltuvused?
* Millised on teenuste kasutamisest saadav kasum või kokkuhoid?
* Kui palju läheb maksma teenuste loomine?
* Milline on Suurandmete süsteemi majandusmudel?

Andmed kui informatsioon omab majanduslikus vaates primaarset tähtust, kuid üldjuhul ei ole võimalik informatsiooni kasutada põllumajandusvaldkonnas ilma täiendava analüüsi ja tootmistegevuse täiustamiseta majandusliku tulu teenimiseks. Kuna Suurandmete süsteem baseerub primaartootmises kogutavatel tegevustel, mis omakorda hõlmavad tootmistehnoloogilisi tegevusi, siis on vajalik kogutavad andmed analüüsida, võimaldada operatiivset andmetel tuginevaid teenuseid ning seeläbi tagada sisend uuesti tootmisprotsesside täpsustamiseks. Selle tõttu on majandusanalüüsi aluseks eelnevalt kaardistatud teenused, mis võimaldaksid põllumajanduses taimekasvatuse tegevusala ettevõtetel tõhustada tootmistegevust. Mitmemõõtmeline majandusanalüüs viidi läbi tootmisteenuste ja mõningatele kaugseire teenustele (vt teenuste valik). Analüüsi kuulusid järgmised teenused:

* Taimekaitse soovitused ja prognoos; Taimekahjustajate kaardistamine; Taimekahjustajate leviku info jagamine ja taimekahjustajate leviku prognoos;
* Huumusbilanss;
* NPK bilanss (integreeritav väetus- ja lupjamissoovitustega);
* E-põlluraamat (minimaalne funktsionaalsus);
* Külvikorra kavandaja;
* Integreeritud agrometeoroloogia andmete jagamise teenus;
* Toetuse taotlejale esitatavate nõuete info jagamine; Toetuste taotluste automaatesitamise teenused;
* Registri teenused: Väetiste registri teenused; Taimekaitsevahendite registri veebiteenused; Sordiregistri veebiteenused;
* FADN andmete visualiseerimise teenus; FADN standardnäitajate jagamise teenus; Majandusliku suuruse ja tootmistüübi kalkulaator (FADN)
* Sordivaliku ja seemne kvaliteedi valiku soovitused; Sordivõrdluskatsete andmete jagamine;
* Liigniiskete alade seire (sisend veeindeksite teenusele);
* Saagikuse prognoos ja kvaliteedi mudel.

Majandusanalüüs viidi läbi sotsiaalmajanduslike mõjude olulisuse hindamise alustel, kuhu kaasati järgmised hindamise komponendid:

**Joonis 18 Sotsiaalmajanduslike mõjude hindamise komponendid**

Eelmärgitud dimensioonide hindamiseks on vajalik teada teenuseid ja kasutajaid ehk sihtrühmi. Käesolevas analüüsi tulem koondati mitmemõõtmelisse tabelisse. Iga teenuse juures hinnati esimese tegevusena osapoolt ning nende seost sihtgrupiga (tootjad). Osapooled olid järgmised:

* Tootja (sihtgrupp);
* Nõustamisasutus;
* Teadusasutus;
* Finantsasutus;
* Riigiasutus;
* Sisendite pakkuja, 3-osapool.
* Teenuse pidaja

Mõjude olulisus avaldub sihtrühma suurusest ja mõju ulatusest. Sihtrühm on suur, kui teenust kasutavate isikute osakaal on üle 50%. Keskmine sihtrühma suurus on, kui kasutajaid on 15-50% ja väike kui alla 15%. Mõju ulatus on suur kui sihtrühma (inimene, ettevõte, keskkond) senine toimimine võib muutuda märkimisväärselt võrreldes varasemaga, sh tekib oluline kasu. Ebasoovitavate mõjudega kaasneb ka risk, mida käesolevas analüüsi ei käsitletud.  Mõju ulatuse hindamiseks kasutati majanduslikke dimensioone nagu teenuse tõhusus ja teenuse mõjusus.

**Joonis 19. Majandusliku tõhususe ja mõjususe hindamise võimalikud dimensioonid ja mõõdikud**

Teenuse tõhususe hindamise mõõdikutena kasutati nii ajalise kui ka rahalise kasu hindamist. Mõõdikute osas sobitusid kokku müügitulu (tulu saagikahju vältimisest), kokkuhoid tootmissisendites (kasu väetiste ja taimekaitsevahendite kokkuhoiust) ja tulu keskkonnahoiust. Juhul kui projekti käigus lisandub teenuseid, siis võib lisanduda ka täiendavaid mõõdikuid.

Majanduslikku kasu hindamiseks arvutati puhastulu järgmiste valemitega:

Otsene tulu=(hinnatud kasu kogus × mõjutatud ühikud (ha, tootja) × maksumus) + (mõjutatud ajaühik (tööaja vähenemine) × keskmine tööjõukulu tunni maksumus)

Otsene kulu=(teenuste andmehõive, töötlus × tõõjõukulu tunni maksumus) + (teenuse arendus × tõõjõukulu tunni maksumus) + (teenuse püsikulu × tõõjõukulu tunni maksumus)

Puhastulu=otsene tulu – otsene kulu.

Teenuse mõjusust hinnati halduskoormuse ja teenuse kvaliteedi ning kättesaadavuse indikaatoritega. Lisaks rakendati PENG mudelis käsitletavat IT eelise astet (nn *Farbey Ladder*). Mida suurem on hinnang süsteemile (antud juhul teenusele), seda keerukam ja kõrgema tasemega võimalusi infosüsteemi teenus pakub. Mõjusust hinnati osapoolte gruppide lõikes skaalal 0-5 punti ja Farbey astet samuti osapoolte lõikes skaalal 1-8 punkti, mis omakorda tähendab, et omistatud punktid lõpptulemuse saamiseks summeeriti.

Hindamiseks kasutati täiendavaid kvantitatiivseid baasandmeid järgmiselt:

* Tootjate arv 2018; Kultuuride pindala 2018 (allikas: Statistikaamet PM0281; PMS108; PRIA toetuste ja põllumassiivide register 2018);
* Keskmine saagikus 2014-2018 (allikas: Statistikaamet PM20);
* Keskmine kokkuostuhind 2016-2019 kv (allikas: Eesti Konjunktuuriinstituut 2019);
* Tööjõukulu tunnis tegevusalade lõikes 2018 (allikas: Statistikaamet PA001);
* Sisendite maksumus kattetulu arvutamise skeemides 2018 (Põllumajandusuuringute Keskus 2018);

**Tulemused**

**Sihtgrupi suurus** - Sihtrühma suurus ja seeläbi mõju ulatus on kõikide kaardistatud teenuste puhul teenuse pidaja vaates üle 50% jäädes koguni vahemikku 80-100%. Tulenevalt osapoolest on sihtgrupi (tootjad) suuruse varieeruvus siiski olemas, kus konservatiivse hinnangu kohaselt nõustamisasutuse huvi võib piirduda kuni 50% ja finantsteenuse pakkuja huvi kuni 60% osakaaluga tootjate üldkogumist. Teadus- ja riigiasutused on ilmslet huvitatud maksimaalsest võimalikust tootjate sihtgrupi osakaalust. Kuna sotsaalmajanduslikus analüüsis loetakse sihtgruppi suureks, kui see ületab 50%, siis järeldub, et valitud teenust ulatus on suur. **Mõjususe kasutegurid** - Mõju ulatuse hindamiseks sotsiaalses vaates omistati konsortsiumi liikmete poolt hinnangud teenuse vajalikkuse ja võimaliku mõjususe kohta. Hindamisel lähtuti teenuste kaardistamisel koostatud kirjeldusest ning võimalikest positiivsetest muutustest. Lisaks määrati IT eelise mõju aste (nn *Farbey* aste). Teenuseid hinnati seotud osapoolte lõikes, mistõttu punktid summeerusid. Mõjususe ja IT eelise astme põhjal sai kõige enam hindepunkte ePõlluraamatu teenus, huumusbilansi ja NPK bilansi kalkulaator koos väetus- ja lupjamissoovitustega ning taimekaitse soovitused ja prognoos, aga samuti sordivõrdluskatsete andmete jagamine ja sordivaliku soovitused. Viimasega võrdväärseid punkte jagas ka saagikuse prognoos ja kvaliteedi mudel.

|  |  |  |  |
| --- | --- | --- | --- |
| Teenus | Hinnang sihtrühma suurusele (3-suur, 2- keskmine, 1-väike) | Hinnang mõjususele (0-35 punkti) | Farbey aste  (7-56 punkti) |
| ePõlluraamat | suur, kuni 95% | **27** | **29** |
| Huumusbilanss | suur, kuni 95% | **18** | **21** |
| NPK bilanss (integreeritav väetus- ja lupjamissoovitustega) | suur, kuni 95% | **18** | **21** |
| Taimekaitse soovitused ja prognoos; Taimekahjustajate kaardistamine; Taimekahjustajate leviku info jagamine ja taimekahjustajate leviku prognoos | suur, kuni 95% | **17** | **21** |
| Sordivõrdluskatsete andmete jagamine; Sordivaliku ja seemne kvaliteedi valiku soovitused | suur, kuni 95% | **17** | **21** |
| Saagikuse prognoos ja kvaliteedi mudel | suur, kuni 95% | **17** | **21** |
| Külvikorra kavandaja | suur, kuni 95% | 15 | 18 |
| Integreeritud agrometeoroloogia andmete jagamise teenus | suur, kuni 95% | 15 | 14 |
| Toetuse taotlejale esitatavate nõuete info jagamine; Toetuste taotluste automaatesitamise teenused | suur, kuni 95% | 14 | 14 |
| Registri teenused: Väetiste registri teenused; Taimekaitsevahendite registri veebiteenused; Sordiregistri veebiteenused (olemas) | suur, kuni 95% | 20 | 14 |
| FADN andmete visualiseerimise teenus (+standartnäitajate jagamine); Majandusliku suuruse ja tootmistüübi kalkulaator (FADN) | suur, kuni 100% | 14 | 14 |
| Liigniiskete alade seire (sisend veeindeksite teenusele (komplektis) | suur, kuni 95% | 12 | 14 |

**Tabel 23. Sihtrühma ulatuse ja majandusliku mõjususe koondtulemused**

Järgnevalt viidi läbi majandusliku tõhususe hindamine vastavalt võimalikele teenuse tulu ja kulu komponentidele. Majandusanalüüsi käigus leitud puhastulu teenustele on arvutatud võrdlemise eesmärgil ja seetõttu ei tohi leitud tulu teenuste lõikes summeeria. Teenused omavad üksteisega sünergiat ja kõikide teenuste rahalise puhastulu summeerimiseks on vajalik koostada täiendavad arvutused.

**Tõhususe kasutegur** - leiti otsese tulu ja otsese kulu hindamisel, mille arvutuslik väljund on puhastulu. Teenuste arenduskulu juures kasutati arvestuslikku meetodit, kus konkreetsel aastal tehtav arenduskulu jagati võimaliku teenuse kasuliku elueaga (keskmiselt 5 aastat), mille tulemusel saadi teada arvestuslik arenduskulu ühe aasta kohta. Järeldame, et võimalik on saavutada suurim puhastulu huumusbilansi ja sordivõrdluskatsete ning sordivaliku ja seemne kvaliteedi valiku soovitustest (vastavalt kuni 24,6 ja 23,0 mln €/a). Järgnevad NPK bilansi ja külvikorra kavandja teenus. Sealjuures on oluline arvestada, et külvikorra kavandaja teenusele on vajalik e-põlluraamatu olemasolu või analoog. E-põlluraamatul on oluline roll nii sihtgrupi kui ka riiklike andmete kogumise seisukohalt. Puhastulu arvutamisel võeti e-põlluraamatu võimalike tulude osas arvesse minimaalne funktsionaalsus, mistõttu ei peegeldu erasektori poolt pakutavate teenuste tulu. Tulude jaotuse põhjal näeme, et e-põlluraamatu minimaalne funktsonaalsus aitab kokku hoida eelkõige tööaega (kuni 2,8 mln €/a). Samal ajal külvikorra kavandja on üks võimalikest teenustest, mis tugineb e-põlluraamatule. ePõlluraamatu rakendamine koos võimalike lisateenustega võib esmaste arvutuste kohaselt luua otsest tulu ligikaudu 30 mln €/a.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Teenus | Otsene tulu, mln €/a | Teenuse kulu kokku, mln €/a | Puhastulu, mln €/a | Puhastulu vahemik |
| Taimekaitse soovitused ja prognoos + Taimekahjustajate kaardistamine; Taimekahjustajate leviku info jagamine ja taimekahjustajate leviku prognoos | kuni 9.3 | kuni 0.08 | kuni 9.3 | 5,0-9,9 mln |
| Huumusbilanss | kuni 23.1 | kuni 0.04 | kuni 23 | >20,0 mln |
| NPK bilanss (integreeritav väetus- ja lupjamissoovitustega) | kuni 17.6 | kuni 0.04 | kuni 17.6 | 15,0-19,9 mln |
| ePõlluraamat | kuni 2.8 | kuni 0.54 | kuni 2.3 | 1,0-4,9 mln |
| Külvikorra kavandaja | kuni 15.1 | kuni 0.11 | kuni 14.99 | 10,0-14,9 mln |
| Integreeritud agrometeoroloogia andmete jagamise teenus | kuni 0.7 | kuni 0.08 | kuni 0.6 | <1,0 mln |
| Toetuse taotlejale esitatavate nõuete info jagamine; Toetuste taotluste automaatesitamise teenused | kuni 2 | kuni 0.59 | kuni 1.4 | 1,0-4,9 mln |
| Registri teenused: Väetiste registri teenused; Taimekaitsevahendite registri veebiteenused; Sordiregistri veebiteenused (olemas) | kuni 0.9 | kuni 0.2 | kuni 0.7 | <1,0 mln |
| FADN andmete visualiseerimise teenus (+standartnäitajate jagamine); Majandusliku suuruse ja tootmistüübi kalkulaator (FADN) | kuni 0.3 | kuni 0.02 | kuni 0.3 | <1,0 mln |
| Sordivõrdluskatsete andmete jagamine; Sordivaliku ja seemne kvaliteedi valiku soovitused | kuni 24.9 | kuni 0.34 | kuni 24.6 | >20,0 mln |
| Saagikuse prognoos ja kvaliteedi mudel | kuni 0.2 | kuni 0.08 | kuni 0.1 | <1,0 mln |
| Liigniiskete alade seire (sisend veeindeksite teenusele (komplektis) | kuni 12.7 | kuni 0.59 | kuni 12.1 | 10,0-14,9 mln |

**Tabel 24. Teenustele hinnatud otsene tulu, kulu ja võimalik puhastulu vahemik**

Komplekse sotsiaalmajandusliku hinnangu andmiseks omistati puhastulu vahemikele punktid <1,0 mln - 10p; 1,0-4,9 mln -15p; 5,0-9,9 mln - 20p; 10,0-14,9 mln - 25p; 15,0-19,9 mln-30p; >20,0 mln - 35p. Joonisel 20 nähtub, et sotsiaalmajanduslikult on e-põlluraamat üks olulisemaid teenuseid (71 punkti), sõltumata sellest, et minimaalne funktsionaalsus ei võimalda saavutada suurt puhastulu.

**Joonis 20. Sotsiaalmajanduslik koondhinnang valitud tootmisteenustele**

Järgnevad teenused, mis võimaldavad vältida saagikahju, suurendada saagikust ja minimeerida tootmisriski. Ajalisest kokkuhoiust sõltuv tulu (joonis) on oluline eelkõige erinevate registriteenuste, nõuete ja agrometeoroloogiliste andmete jagamise teenuste juures. Samas on tegemist suhtelise näitajaga, kuna kogu ajakulu kasutust ei ole põllumajandusvaldkonnas hinnatud. Siinkohal on aluseks ajakulu vähenemine (nt nõuete otsimisel ja lugemisel), mille kogumaht ilma IT teenuseta ei ole täpselt teada. IT teenus aitab vähendada ajakulu, kuid ei muuda seda olematuks.

# Suurandmete süsteemi loomise teekaart

|  |  |  |  |
| --- | --- | --- | --- |
| Task Name | Duration | Start | Finish |
| **1. Suurandmete süsteem\_0\_etapp / Eeltegevused (0-6 kuud)** | **173 days** | **Mon 09.09.19** | **Fri 28.02.20** |
| **Teadmussiirde projekti II etapp (eeltegevused)** | **173 days** | **Mon 09.09.19** | **Fri 28.02.20** |
| Teadmussiirde projekti I etapi järeldused on tehtud | 14 days | Mon 09.09.19 | Sun 22.09.19 |
| **Valikud Suurandmete süsteemi II etapi hanke korraldamiseks on tehtud** | **14 days** | **Sun 22.09.19** | **Sat 05.10.19** |
| Kokku on lepitud Suurandmete süsteemi ahitektuurne struktuur | 7 days | Sun 22.09.19 | Sat 28.09.19 |
| Kokku on lepitud realiseeritavate teenuste täpsustatud eelvalik | 14 days | Sun 22.09.19 | Sat 05.10.19 |
| Kokku on lepitud teenused mida võib osutada ka 3-s sektor | 14 days? | Sun 22.09.19 | Sat 05.10.19 |
| Kokku on lepitud juriidilised 0\_etapi juriidilised valikud | 14 days | Sun 22.09.19 | Sat 05.10.19 |
| **Alternatiivsete teenusehaldajate (3-s sektor) võimaluste hindamine** | **30 days** | **Sun 06.10.19** | **Mon 04.11.19** |
| **Suurandmete süsteemi II etapi hange** | **146 days** | **Sun 06.10.19** | **Fri 28.02.20** |
| Hanke korraldaja koostab II etapi lähteülesande eeldatava skoobi | 30 days | Sun 06.10.19 | Mon 04.11.19 |
| Hanketegevuse läbiviimine | 88 days | Tue 05.11.19 | Fri 31.01.20 |
| Hange on õnnestunud, hanget ei vaidlustata ja hankeleping on ettevalmistatud | 28 days | Sat 01.02.20 | Fri 28.02.20 |
| **Juriidilised tegevused (0\_etapp)** | **160 days** | **Sun 22.09.19** | **Fri 28.02.20** |
| Juriidilise toe tagamine lähtuvalt I etapi järeldustest ja sisust hanke läbiviimisel | 160 days | Sun 22.09.19 | Fri 28.02.20 |
| Seaduseelnõu väljatöötamise kavatsuse põhimõtted on välja töötatud | 160 days | Sun 22.09.19 | Fri 28.02.20 |
| **2. Suurandmete süsteem\_I etapp / Põhitegevused (6 kuud - 2 aastat)** | **550 days** | **Sat 29.02.20** | **Tue 31.08.21** |
| **Teadmussiirde projekti II etapp (teostus)** | **550 days** | **Sat 29.02.20** | **Tue 31.08.21** |
| Hange on edukaks kuulutatud ja hankeleping sõlmitud | 1 day | Sat 29.02.20 | Sat 29.02.20 |
| **Suurandmete süsteemi haldamise organsiatsioon** | **549 days** | **Sun 01.03.20** | **Tue 31.08.21** |
| Paika pannakse organisatsioon, rollid, ülesanded | 7 days | Sun 01.03.20 | Sat 07.03.20 |
| Olemasolevate andmekogude vastutatvad töötlejad on kaasatud | 7 days | Sun 01.03.20 | Sat 07.03.20 |
| Juhtimine, arendusohje, jätkusuutlikkuse tagamine | 542 days | Sun 08.03.20 | Tue 31.08.21 |
| Olemasolevate teenuste ajakohasuse tagamine ja uute teenuste teke | 150 days | Sun 08.03.20 | Tue 04.08.20 |
| **Suurandmete baassüsteemi arendus vastavalt hanketingimustes valitud arhitektuurile** | **180 days** | **Sun 01.03.20** | **Thu 27.08.20** |
| **Keskkondade loomine** | **180 days** | **Sun 01.03.20** | **Thu 27.08.20** |
| Andmehõive keskkonna loomine (sõltub valitud arh. Lahendusest) | 180 days | Sun 01.03.20 | Thu 27.08.20 |
| Andmetöötluskeskkonna loomine (sõltub valitud arh. Lahendusest) | 180 days | Sun 01.03.20 | Thu 27.08.20 |
| Publitseerimiskeskkonna loomine(sõltub valitud arh. Lahendusest) | 180 days | Sun 01.03.20 | Thu 27.08.20 |
| Andmepakettide koostamine | 180 days | Sun 01.03.20 | Thu 27.08.20 |
| Analüütika ja masinõppe keskkonna loomine | 180 days | Sun 01.03.20 | Thu 27.08.20 |
| **Põhiandmete haldus** | **180 days** | **Sun 01.03.20** | **Thu 27.08.20** |
| Metaandmete haldus | 180 days | Sun 01.03.20 | Thu 27.08.20 |
| Klassifikaatorite ja koodilistide haldus | 180 days | Sun 01.03.20 | Thu 27.08.20 |
| **Baasteenuste loomine** | **180 days** | **Sun 01.03.20** | **Thu 27.08.20** |
| Andmebaasi loomine | 180 days | Sun 01.03.20 | Thu 27.08.20 |
| Veebiteenused | 180 days | Sun 01.03.20 | Thu 27.08.20 |
| Kastajaliidesed | 180 days | Sun 01.03.20 | Thu 27.08.20 |
| Kasutajateenused | 180 days | Sun 01.03.20 | Thu 27.08.20 |
| **Teenuste skoobi realiseerimine vastavalt hankes kajastatud eelarve mahule** | **453 days** | **Sun 01.03.20** | **Thu 27.05.21** |
| **Lähtuvalt majandusanalüüsi tulemustest ja hankes fikseeritud mahuhinnangutest algab teenuse realiseerimine** | **1 day** | **Sun 01.03.20** | **Sun 01.03.20** |
| Teenus 1 / Taimekaitse soovitused ja prognoos + Taimekahjustajate kaardistamine; Taimekahjustajate leviku info jagamine ja taimekahjustajate leviku prognoos | 1 day | Sun 01.03.20 | Sun 01.03.20 |
| Teenus 2 / Huumusbilanss | 1 day | Sun 01.03.20 | Sun 01.03.20 |
| Teenus 3 / NPK bilanss (integreeritav väetus- ja lupjamissoovitustega) | 1 day | Sun 01.03.20 | Sun 01.03.20 |
| Teenus 4 / Integreeritud agrometeoroloogia andmete jagamise teenus | 1 day | Sun 01.03.20 | Sun 01.03.20 |
| **Andmekogude prioriteetsuse valik (Skoobi realiseerimise järjekord)** | **30 days** | **Sun 01.03.20** | **Mon 30.03.20** |
| Sisendite seoste ülevaade andmekogu v andekogude grupi lõikes - kõige enam teenustega seoseid omav andmekogu suunatakse töösse kõige varem | 5 days | Sun 01.03.20 | Thu 05.03.20 |
| Andmekogude omanikega koostööformaadi kokkulepe | 7 days | Fri 06.03.20 | Thu 12.03.20 |
| Kõikide ligipääsude korraldustegevus | 30 days | Sun 01.03.20 | Mon 30.03.20 |
| **Töö andmekogudega** | **439 days** | **Sun 15.03.20** | **Thu 27.05.21** |
| **Valitud teenustega seotud andmekogude tegevused** | **137 days** | **Sun 15.03.20** | **Wed 29.07.20** |
| Teadmussiirde I etapi analüüsis esiletoodud puuduste elimineerimine | 60 days | Sun 15.03.20 | Wed 13.05.20 |
| Andmekogude liidestusvajaduse standardikohane realiseerimine | 90 days | Fri 01.05.20 | Wed 29.07.20 |
| **Andmekogude teenusesse integratsioon (teenused 1-4)** | **300 days** | **Sat 01.08.20** | **Thu 27.05.21** |
| T1:Kasutajaliides: Taimekaitse soovitused ja prognoos + Taimekahjustajate kaardistamine | 300 days | Sat 01.08.20 | Thu 27.05.21 |
| T1:Kasutajateenus: Taimekaitse soovitused ja prognoos + Taimekahjustajate kaardistamine | 300 days | Sat 01.08.20 | Thu 27.05.21 |
| T1:Veebiteenus: Taimekaitse soovitused ja prognoos + Taimekahjustajate kaardistamine | 300 days | Sat 01.08.20 | Thu 27.05.21 |
| T1:Andmebaas: Taimekaitse soovitused ja prognoos + Taimekahjustajate kaardistamine | 300 days | Sat 01.08.20 | Thu 27.05.21 |
| T2-T4 liidestused (Huumusbilanss, NPK bilanss, Integreeritud agrometeoroloogia) | 300 days | Sat 01.08.20 | Thu 27.05.21 |
| **Teostatud teenuse tutvustustegevused** | **122 days** | **Thu 01.04.21** | **Sat 31.07.21** |
| Sihtgrupile suunatud infopäevad | 21 days | Tue 01.06.21 | Mon 21.06.21 |
| Sihtgrupile suunatud koolitustegevus | 40 days | Tue 22.06.21 | Sat 31.07.21 |
| Kirjalikud materjalid ja publikatsioonid | 122 days | Thu 01.04.21 | Sat 31.07.21 |
| **Juriidilised tegevused (I etapp)** | **180 days** | **Sun 01.03.20** | **Thu 27.08.20** |
| Juhtrüma töökord on välja töötatud ja liikmed on määratud | 30 days | Sun 01.03.20 | Mon 30.03.20 |
| **Välja on töötatud suurandmete süsteemi üldised põhimõtted** | **180 days** | **Sun 01.03.20** | **Thu 27.08.20** |
| Suurandmete süsteemi tegevuse üldpõhimõtted, katusmõisted, vastutuse piirid, andmete jagamise reeglid ja ülesannete delegeerimine erasektorile | 180 days | Sun 01.03.20 | Thu 27.08.20 |
| **Analüüsitud on suurandmte süsteemi reguleerimise võimalusi ja välja on töötatud suurandmete süsteemi keskse tehnilise lahenduse õiguslikud valikud ja muudatusvajadused õigusaktides** | **180 days** | **Sun 01.03.20** | **Thu 27.08.20** |
| Analüüsitud on seaduslikku alust andmete töötlemiseks, kirjeldatud on andmekogu pidamise eesmärk, otsustatud on vastutavad ja volitatud töötleja ja nende ülesanded | 180 days | Sun 01.03.20 | Thu 27.08.20 |
| Analüüsitud on ülesannete delegeerimise võimalused eraõiguslikele isikutele | 160 days | Sat 21.03.20 | Thu 27.08.20 |
| Välja on töötatud andmete kaitse põhimõtted | 180 days | Sun 01.03.20 | Thu 27.08.20 |
| Kirjeldatud on suurandmete süsteemi andmekoosseis | 180 days | Sun 01.03.20 | Thu 27.08.20 |
| Kirjeldatud on kohustuslikud andmeandjad ja info edastamise viis | 180 days | Sun 01.03.20 | Thu 27.08.20 |
| Välja on töötatud andmete väljastamise ja juurdepääsu põhimõtted | 180 days | Sun 01.03.20 | Thu 27.08.20 |
| Välja on töötatud on õigusaktide muudatusvajadused teenuste realiseerimiseks | 180 days | Sun 01.03.20 | Thu 27.08.20 |
| **Eelnõu väljatöötamine** | **365 days** | **Wed 01.07.20** | **Wed 30.06.21** |
| **Eelnõu kooskõlastamine ministeeriumitega** | **63 days** | **Wed 01.07.20** | **Tue 01.09.20** |
| Eelnõu kooskõlastamien justiitsministeeriumiga | 30 days | **Tue 01.09.20** | Wed 30.09.20 |
| Eelnõu esitamine vabariigi valitsusele | 1 day | Thu 15.10.20 | Thu 15.10.20 |
| Eelnõu menetlus riigikogus | 137 days | Fri 16.10.20 | Mon 01.03.21 |
| Rakendusaktide väljatöötamine | 122 days | Mon 01.03.21 | Wed 30.06.21 |
| **Suurandmete süsteem\_II\_etapp / Põhitegevused (2aastat - 5 aastat)** | **1095 days** | **Wed 01.09.21** | **Fri 30.08.24** |
| **Teadmussiirde projekti II etapp (järeltegevused)** | **1095 days** | **Wed 01.09.21** | **Fri 30.08.24** |
| **Varem realiseerimata teenuste võimalik liitumine suurandmete süsteemiga** | **1095 days** | **Wed 01.09.21** | **Fri 30.08.24** |
| **Automaatliitumise võimalused (I etapi andmekogude analüüsi põhjal)** | **1095 days** | **Wed 01.09.21** | **Fri 30.08.24** |
| Suurandmete süsteemi struktuur ja liidestusvalmidus on olemas | 1 day | **Wed 01.09.21** | Wed 01.09.21 |
| Järgides 0\_etapis kehtestatud liitumisstandardit toimub uute teenuste eeldusteks vajalike andmekogude liitumine suurandmete süsteemiga minimaalse ajakuluga | 1095 days | Wed 01.09.21 | Fri 30.08.24 |
| **III osapooltele suunatud tegevused ja III sektori teenuste teke** | **180 days** | **Wed 01.09.21** | **Sun 27.02.22** |
| Teenuste omanike korraldatud võimalike uute teenuste sihtgrupile suunatud tutvustustegevus | 21 days | Wed 01.09.21 | Tue 21.09.21 |
| Täiendavate rahastusvõimaluste leidmine uute teenuste tekkeks | 60 days | Wed 01.09.21 | Sat 30.10.21 |
| II etapis teostamata kuid sisuliselt vajalike teenuste realiseerimine vastavalt eelarvele | 180 days | Wed 01.09.21 | Sun 27.02.22 |
| **Teenuste ajakohasuse tagamine ja arendustegevus** | **1095 days** | **Wed 01.09.21** | **Fri 30.08.24** |
| **Olemasolevate teenuste monitooring** | **1095 days** | **Wed 01.09.21** | **Fri 30.08.24** |
| Tagasiside allikate monitooring | 1095 days | Wed 01.09.21 | Fri 30.08.24 |
| Andmete analüüs ja süntees | 1095 days | Wed 01.09.21 | Fri 30.08.24 |
| Parandusettepanekute tegemine teenustesse | 1095 days | Wed 01.09.21 | Fri 30.08.24 |
| **Uute võimalike teenuste monitooring** | **1095 days** | **Wed 01.09.21** | **Fri 30.08.24** |
| Tagasiside monitooring | 1095 days | Wed 01.09.21 | Fri 30.08.24 |
| Andmete analüüs ja süstees | 1095 days | Wed 01.09.21 | Fri 30.08.24 |
| Uute teenusvõimaluste kaardistamine | 1095 days | Wed 01.09.21 | Fri 30.08.24 |
| Uute teadlusprojektide sisendite kaardistamine | 1095 days | Wed 01.09.21 | Fri 30.08.24 |
| Huvigruppidega suhtlemine potentsiaalsete teenuste kontekstis | 1095 days | Wed 01.09.21 | Fri 30.08.24 |
| **Teenuse eksportvõimalused** | **1095 days** | **Wed 01.09.21** | **Fri 30.08.24** |
| Teenuste tutvustusvõimaluste loomine ja ettevalmistused | 1095 days | Wed 01.09.21 | Fri 30.08.24 |
| Teenuse unikaalsuse monitooring EU piires | 1095 days | Wed 01.09.21 | Fri 30.08.24 |
| **Juriidiline tegevuskava (II etapp)** | **1095 days** | **Wed 01.09.21** | **Fri 30.08.24** |
| Suurandmete süsteemi I etapi tulemi juriidiline analüüs ja vajadusel õigusaktide muudatuste väljatöötamine | 1095 days? | **Wed 01.09.21** | **Fri 30.08.24** |
| **II tegevusetapp on lõpetatud ja suurandmete süsteem on rakendatud** | **1 day** | **Sat 31.08.24** | **Sat 31.08.24** |

# Viited

[1] Elcano: A Geospatial Big Data Processing System based on SparkSQL, Jonathan Engélinus and Thierry Badard Centre for Research in Geomatics (CRG), Laval University, Québec, Canada 2018 - <https://www.scitepress.org/Papers/2018/67946/pdf/index.html>

[2] Statistikaameti tegevuskava Eesti andmehalduse juhtimisel - <https://www.stat.ee/dokumendid/1413034>

[3] DDI standard - <https://www.ddialliance.org/>

[4] RIA andmekvaliteediuuringu lõpparuanne - <https://www.ria.ee/sites/default/files/content-editors/publikatsioonid/andmekvaliteedi_uuringu_lopparuanne.pdf>

[5] Open Data standard - <https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=odata>

[6] Geomeetria andmetüübi kirjeldus - <https://www.postgresql.org/docs/9.4/datatype-geometric.html>