TIEDON LAATU

VIRPI VIRTANEN S-PANKKI JA JANNE ERÄNEN PERIGEUM

S-Pankki

SISÄLTÖ

- Miksi tiedon laatua pitää tarkkailla?
- Sanktioitu ympäristö
- Mitä tiedosta voidaan tarkkailla?
  - Numeeriset ja merkkimuotoiset tarkastelut
- DQ engine
  - Tekniset speksit
- Numeeriset ja merkkimuotoiset hälyt
- Louhintaprosessin tulokset
- Ideoita ja tarpeita

S-Pankki

MIKSI TIEDON LAATUA PITÄÄ TARKKAILLA

- Virheellistä tai puuttuvaa dataa ei voi käyttää tiedon lähteenä.
- Tiedon on oltava todistetusti oikeaa, täsmäävää ja ristiin tarkistettua tai tarkistettavaa.
- Viranomaisvaade sanktioineen asettaa konkreettiset määritykset tiedon laadulle.

S-Pankki

SANKTIOITU YMPÄRISTÖ

- Sanktiot vääristä luvuista ja viivästyneistä raporteista, maineen menetys ja uudelleen tehtävien raporttien työmäärä
- Rikemaksu-tyyppisen sanktion suuruus on 5000 -100 000 euroa ja Finanssivalvonta on määrännyt näitä jo raporttien myöhästyessä 3 pankkipäivää
- Seuraamismaksu-tyyppisen sanktion suuruus voi olla 1-5 miljoonaa euroa. Sanktiota ei välttämättä määrätä jos korjaukset tehdään oma-aloitteisesti virheen havaitsemisen jälkeen
- Rikemaksu ja seuraamismaksu ovat laissa määriteltyjä maksuja.

S-Pankki

FINANSSIVALVONNAN TUTKINTAPYYNNÖT

Sanktiot, tutkintapyynnöt ja väärinkäyttöepäilyt

Finanssivalvonnan tutkintapyynnöt ja hallinnolliset seuraamukset 2005–2015

	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
Tutkintapyyntö poliisille	1	3	2	5	-	3	4	5	4	3	5
Julkinen huomautus	2	-	1	1	5	3	3	3	-	-	2
Julkinen varoitus	-	-	-	1	-	-	-	3	2	2	-
Rikemaksu	-	-	-	14*	-	1	5	14	6	2	19
Seuraamusmaksu	-	-	-	-	-	-	-	-	-	1	-

** Luku korjattu 1.10.2013*

- Seitsemälle yritykselle myöhästymisestä rikemaksut marraskuussa 2015 mm. Handelsbanken Finans Abp J raportti 7 pankkipäivää myöhässä, Helsinki Capital Partners Rahastoyhtiö Oy FINREP 3 pankkipäivää myöhässä ja COREP 4 pankkipäivää myöhässä.

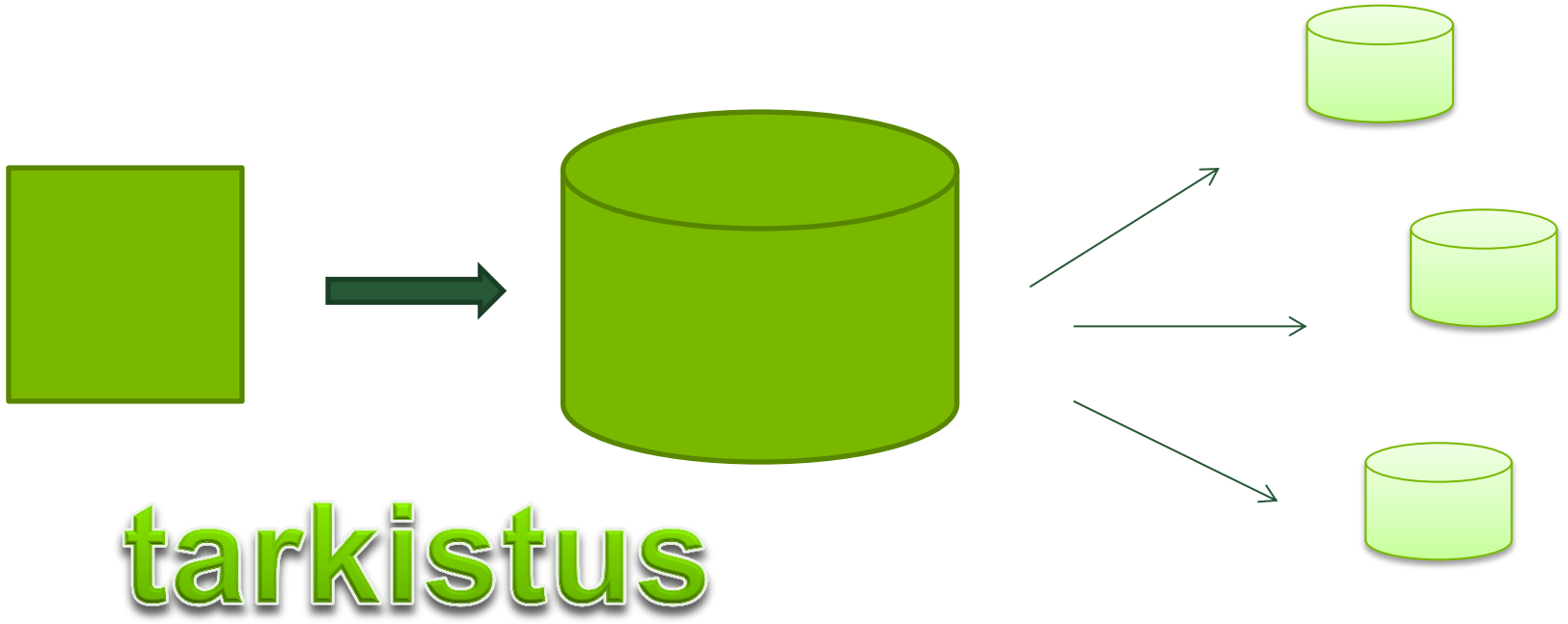
S-Pankki

HAASTEET TARKKAILUSSA

- Tieto voi olla koko aineiston kattava päivittäin tai data voi sisältää vain muutoksen.
  - Koko aineiston käyttäytyminen on harmonisempaa, kuin vain muutosten käyttäytyminen. Hälytysrajat voivat olla tiukempia ja käyttäytyminen ennakoitavampaa.
  - Muutostiedon tarkkailu vaatii oman mittaristonsa, jotta dataperäiset virheet tulee oikein poimituiksi.
- Tarkkailu on pyritty keskittämään alueille, jossa data on mahdollisimman lyhyen ajankohdan tietoa (päivä / kuukausi) ja otos on mahdollisimman suuri.

S-Pankki

TIEDON TARKISTAMINEN



S-Pankki

SAAPUVAN TIEDON TARKASTAMINEN

- Ennen kuin tietoa luetaan sisään tietovarastoon sille voidaan tehdä tarkistuksia
  - Onko tieto saapunut silloin, kuin sen pitää saapua: päivittäin / viikoittain / kuukausittain
  - Onko tieto saapunut kokonaan vain osittain vai kenties tyhjänä
  - Onko tiedon sisällössä tapahtunut suuria muutoksia
    - Miten tietoa arvioidaan?
    - Millaiset muutokset ovat sallittuja?
    - Mikä on aikaikkuna, jossa joudutaan toimimaan?

S-Pankki

DQ ENGINE

- Numeropohjaisen tiedon louhiminen voidaan toteuttaa yksinkertaisesti.
- Tarkastellaan perustilastoarvoja, joiden muutoksesta voidaan päätellä mitä datassa tapahtuu
- Nopeasti data-vaiheella laskettavia asioita ovat
  - Minimit, maksimit, keskiarvot, puuttuvien arvojen määrät, summat, keskiarvot puuttuva arvo nolla-arvona
  - Merkkimuotoisen datan validiussääntöjen tarkasteleminen

S-Pankki

MITEN TARKKAILLAAN – PARAMETRINA KIRJASTO JA TARKKAILUAIKA

```
proc sql noprint;
        create table taulut
        as select LIBNAME, MEMNAME, NOBS, FILESIZE, MODATE from dictionary.tables where libname ="&LIBRARY" and
         datdif(datepart(modate),today(), 'ACT/ACT')<=&datedif;
        select count(memname) into :lkm from taulut
        where datdif(datepart(modate),today(), 'ACT/ACT')<=&datedif;
        %let lkm=&lkm;
        select libname, memname, modate, nobs format=12., filesize format=24. into :lib1-:lib&lkm, :mem1-:mem&lkm, :date1-:date&lkm, :nobs1-:nobs&lkm, :fsize1-:fsize&lkm from taulut
        where datdif(today(),datepart(modate), 'ACT/ACT')<=&datedif;
quit;
```

S-Pankki

DQ ENGINE
ERILLINEN KÄSITTELY NUM JA CHAR

```
%do i = 1 %to &lkm;
proc sql;


        create table char&i as
        select * from dictionary.columns
        where libname="&&lib&i" and memname="&&mem&i" and type='char';

        create table num&i as
        select * from dictionary.columns
        where libname="&&lib&i" and memname="&&mem&i" and type='num';
quit;
%end;
```

S-Pankki

NUMEERISET TARKKAILUT

```
data d_%substr(&&mem&j.,1,30);
                set &&lib&j...&&mem&j end=loppu;
                length var $32.;
                retain %do t=1 %to &&lkmn&j;
                                avg&&var&t max&&var&t  min&&var&t  sum&&var&t nmi&&var&t
                %end; ;
                keep var avg sum min max nmiss;
                %do c=1 %to &&lkmn&j;
                                sum&&var&c+&&var&c;
                                min&&var&c= min(&&var&c, min&&var&c);
                                max&&var&c=max(&&var&c,max&&var&c);
                                if &&var&c = . then nmi&&var&c+1;
                %end;
if loppu;
                %do s=1 %to &&lkmn&j;
                                avg&&var&s=sum&&var&s/_n_;
                %end;
```

S-Pankki

NUMEERISET TARKKAILUT

```
%do l=1 %to &&lkmn&j;
if index("&&lvar&l",'_ID') = 0 and index("&&lvar&l",'_DTTM') = 0 and  index("&&lvar&l",'_RK') = 0  and index("&&lvar&l",'_RUNDATE') = 0
and index("&&lvar&l",'_DT') = 0 and index("&&lvar&l",'_DATE') = 0
then do;

                    var="&&lvar&l";
                    avg=avg&&var&l;
                    max=max&&var&l;
                    min=min&&var&l;
                    sum=sum&&var&l;
                    nmiss=nmi&&var&l;
                    output;
end;
%end;

run;
```

S-Pankki

TARKISTUSTEN AUTOMATISOINTI

- Nyt on dataa!!!
  - Numeeriset parametrit (sum, avg, max jne) kerätään kaikilta tauluilta kaikista kentistä joka päivä
- Mutta mitä sillä tehdään?
  - Haasteet:
    - Millaiset muutokset ovat ok? Millaiset eivät? Onko joskus muuttumattomuus syy huolestua?
    - Tarvitseeko kukin kenttä ja parametri omat sääntönsä?
    - Dataa liikaa säännölliseen visuaaliseen tai ad-hoc tarkasteluun
  - Ratkaisu:
    - Automatisoidaan tarkistukset, mutta millä säännöillä?
      - Logiikka, oma kokemus ja liiketoiminta apuun

S-Pankki

AUTOMATISOINTI

- Tietojen kerääminen ei ole ongelma, tietojen esittäminen ei ole ongelma.
- Tietojen esittäminen siten, että sen perusteella voidaan toimia on ongelma.
  - Missä, miten ja kenelle voidaan näyttää tietoa?
  - Kuoleeko tiedon saaja informaatioähkyyn selatessaan päivittäin tuhansia graafeja? (itseasiassa 5200 kuvaa / päivä) Missä ovat inhimillisyyden rajat?
  - Mikä tieto on tärkeää…? Tuskin rundate, vai onko sittenkin?

S-Pankki

TARKISTUSTEN AUTOMATISOINTI

- Automatisoinnin loogisia haasteita
  - Mikä on ”normaali” vaihteluväli?
  - Onko muutos yleensä edes sallittua?
  - Onko muutoksen suunnan oltava aina sama?
  - Poikkileikkaus- vai muutostiedosto?
    - Poikkileikkauksissa pienempiä prosentuaalisia muutoksia kuin muutostiedostossa -> tiukemmat rajat
    - Muutostiedostossa nollariviset huomattavasti todennäköisempiä kuin poikkileikkaustiedostossa
  - Muuttujaa kuvaavan ilmiön luonne. Onko tapahtumia tasaisesti vai keskittyvätkö ne tiettyyn vaiheeseen kuuta?
    - Palkat tulevat ja eräpäivät painottuvat tiettyyn aikaan kuuta -> vaihtelua luonnostaan, mutta kuinka paljon?

→ Joka taululla ja kentällä oma käyttäytymisensä, omat säännöt!

S-Pankki

TARKISTUSTEN AUTOMATISOINTI

- Ohjaustaulun avulla joustava ja skaalautuva ratkaisu
  - Ohjaustaulun rakenne yhtenevä kerätyn numeerisen datan kanssa
  - Ohjataan tarkistukset keskeisiin numeerisiin kenttiin liiketoiminnan tai oman kokemuksen ohjaamana

Kirjasto	Taulu	Muuttuja	Tunnusluku	Hälyraja
Lib1	Taulu1	Euroja1	Avg	>10
Lib1	Taulu1	Aikaleima	Avg	<=0
Lib2	Taulu2	Euroja2	Max	>5

- Poimitaan numeromassasta ohjaustaulussa mainitut kirjasto-taulu-kenttä-tunnusluku –kombinaatiot kahdelta edeltävältä päivältä ja lasketaan prosentuaalinen ero

S-Pankki

TARKISTUSTEN AUTOMATISOINTI

Säännöt:

Kirjasto	Taulu	Muuttuja	Tunnusluku	Hälyraja
Lib1	Taulu1	Euroja1	Avg	>10
Lib1	Taulu1	Aikaleima	Avg	<=0
Lib2	Taulu2	Euroja2	Max	>5

Makrotetaan raja-arvot (>10, <=0, >5), luupataan kullekin, ja nostetaan ”hälytys” -täppä jos ero rajojen ulkopuolella:

- Set halytys=1 where hälyraja="&&raja&i" and pros_ero &&raja&i))

Havainnot:

Kirjasto	Taulu	Muuttuja	Tunnusluku	Hälyraja	Arvo_eilen	Arvo_tänään	Pros_ero	Hälytys
Lib1	Taulu1	Euroja1	Avg	>10	100 000	107 000	7	
Lib1	Taulu1	Aikaleima	Avg	<=0	14mar2016	14mar2016	0	1
Lib2	Taulu2	Euroja2	Max	>5	100 000	107 000	7	1

S-Pankki

TARKISTUSTEN AUTOMATISOINTI

- Suspektin löydyttyä paljon vaihtoehtoja
  - VA-rapsalle lista suspekteista
  - Mailia valikoidulle jakelulle
    - Eri jakelut eri tauluille/kentille
  - Kerätään pidempi historia ja heitetään graafi VA:han
  - Jne jne

S-Pankki

LOUHINTAPROSESSIN TULOKSET

- Tulosten näyttäminen graafeina on ’turvallisin’ vaihtoehto.
- Graafit kertovat datan muutoksesta ajan suhteen.
- Pienellä koodilla voidaan toteuttaa tuhansia sivuja kuvia, joista voidaan nähdä datassa tapahtuvat poikkeamat.

S-Pankki

LOUHINTAPROSESSIN TULOKSET

```
%macro kuva;
%do l = 2 %to 2;
%let lib=%scan(&libs,&l,*);
%do t = 1 %to &&&lib.lkm ;
%let tablevar=%scan(&&&lib.tablevar,&t,*);
%do v= 1 %to 5;
%let value=%scan(&values,&v,*);
%let table=%scan(&tablevar,1,.);
%let var=%scan(&tablevar,2,.);
PROC SORT
                DATA=WORK.COMPARET_&lib (WHERE=(table = "&table" AND var = "&var") KEEP=modate_c &value table
var)
                OUT=WORK.SORTTempTableSorted
                ;
                BY modate_c;
        RUN;
```

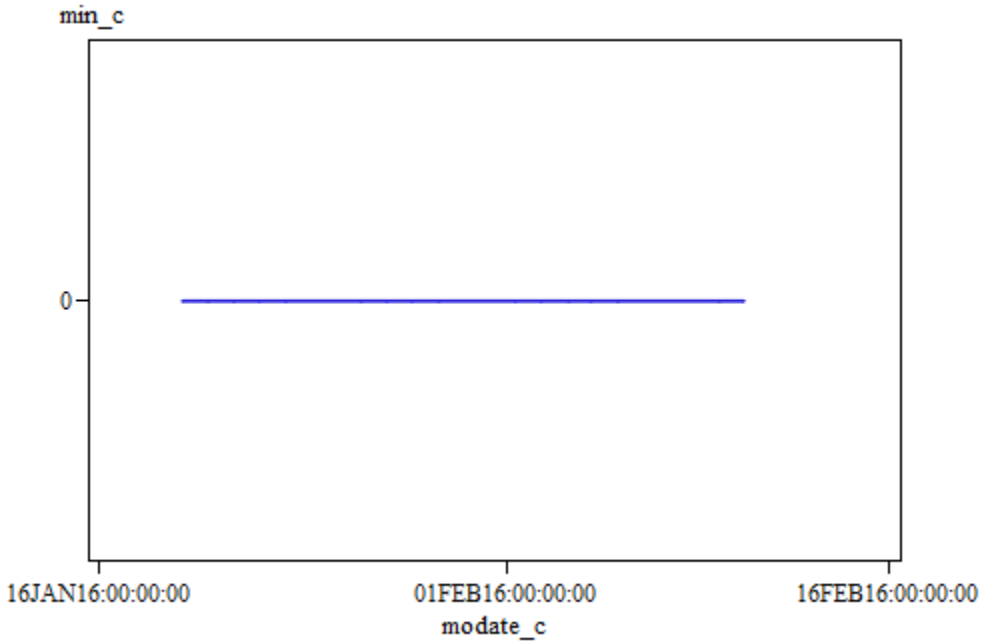
S-Pankki

LOUHINTAPROSESSIN TULOKSET

```
 TITLE1 "Library &lib Table &table Var &var Value &value";
        FOOTNOTE;
PROC GPLOT DATA = WORK.SORTTempTableSorted
        ;
        PLOT &value * modate_c  /
                VAXIS=AXIS1

                HAXIS=AXIS2

        FRAME;
                                format &value nlnum24.;
        RUN;

quit;
%end; %end; %end;              %mend;
                                %kuva;
```

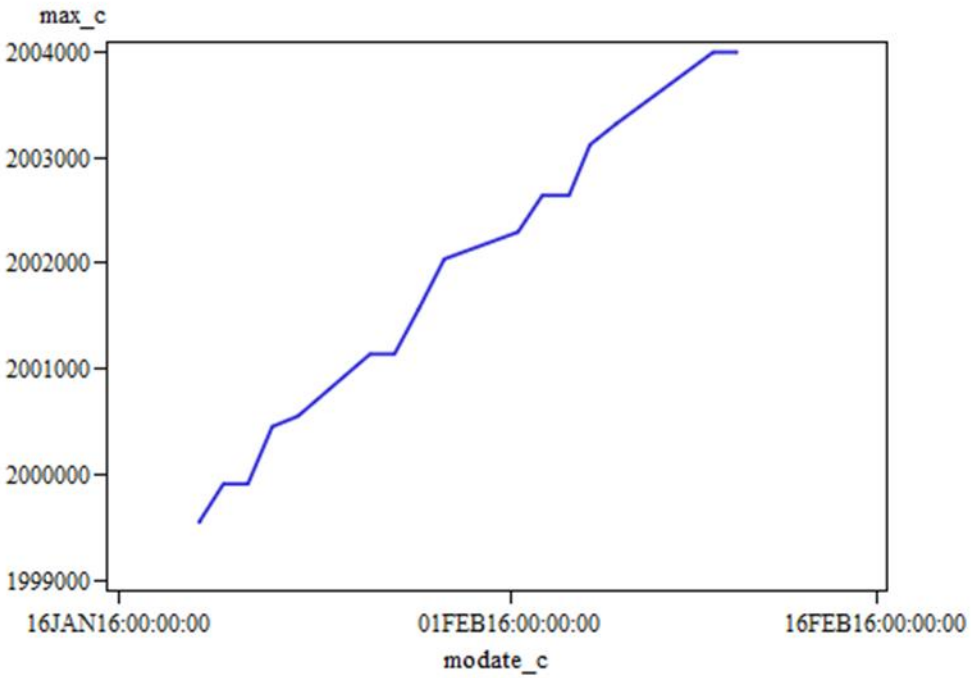
S-Pankki

LOUHINTAPROSESSIN TULOKSET



S-Pankki

LOUHINTAPROSESSIN TULOKSET



S-Pankki

CHAR-KENTTIEN TARKASTUKSET

- Tunnuslukuhistorian (max, avg, sum jne) ei vaihtoehto
- Solutason massa käytävä läpi, mutta miten?
- Mahdollisia tarkistustyyppejä:
  - Sallittu rakenne tunnettu, esim.
    - Kotimainen puhelinnumero
      - alku +358 jonka jälkeen vähintään 6 numeroa. Ei kirjaimia
    - Y-tunnus
      - Alussa 7 numeroa, sitten viiva ja 1 numero
  - Sallittu arvojoukko tunnettu, esim.
    - Firman portfoliossa tuotteet ”012”, ”013”, ja ”014”
      - Löytyvät tuotetaulusta

S-Pankki

CHAR-KENTTIEN TARKASTUKSET

- Sama ohjaustaulurakenne kuin numeerisillakin

Säännöt:

Kirjasto	Taulu	Muuttuja	Tarkistustyyppi	Parametri1
Lib1	Taulu1	Var2	Rakenne	Puhelin
Lib1	Taulu2	Var6	Arvojoukko	Libx.tuotetaulu.tuotekoodi
Lib2	Taulu3	Var8	Arvojoukko	Libx.tuotetaulu.tuotekoodi
Lib2	Taulu3	Var9	Rakenne	Y-tunnus

S-Pankki

CHAR-KENTTIEN TARKASTUKSET

- Kerätään tarkastettavat yhteen

```
%do i=1 %to &lkm.;
   %put Haetaan arvot kirjaston &&kirjasto&i taulun &&taulu&i
        kentästä &&mja&i tarkistusta "&&tarkistus&i" varten.;

   data poiminta&i;
      length kirjasto $8. taulu kentta tarkistus parametri $100. arvo $200.;
      set &&kirjasto&i..&&taulu&i;
      arvo=&&mja&i;
      kentta="&&mja&i";
      taulu="&&taulu&i";
      kirjasto="&&kirjasto&i";
      tarkistus="&&tarkistus&i";
      parametri="&&parametri&i";
      keep kirjasto taulu kentta tarkistus parametri arvo;
   run;

   proc append base=kaikki_yhteen data=poiminta&i; run;

%end;
```

Taulu1		
Var1	Var2	Var3
A	+358401234567	
B	+35850ÖÖÖ	

Taulu2		
Var4	Var5	Var6
a	6	012
b	7	013

Taulu3		
Var7	Var8	Var9
åå	013	1234567-8
öö	ÖÖÖ	123-45678

T1	Var2	+358401234567
T1	Var2	+35850ÖÖÖ
T2	Var6	012
T2	Var6	013
T3	Var8	013
T3	Var8	ÖÖÖ
T3	Var9	1234567-8
T3	Var9	123-45678

- Etuja:
  - Arvot yhdessä kentässä
  - Muoto sama kuin ohjaustaulussa
  - Vain yksi raskas arvojoukkovertailu per taulu

S-Pankki

CHAR RAKENNETARKASTUKSET

- Perl-regular expressioneilla
  - Vähemmän bugialtis kuin perinteisempi substr-scan-sekamelska
  - Helpompi ylläpitää
  - Tehokkaampi
  - Kaksiosainen rakenne data stepin sisällä:
    - Luodaan sääntö (prxparse)
    - Kutsutaan sääntöä (prxmatch)
- Esimerkkejä:
  - p_numero = prxparse("/\d+/"); /* sääntö: sisältää numeroita */
  - If prxmatch(p_numero,arvo)>0 then numeroita=1; /* tarkistus */

S-Pankki

CHAR RAKENNETARKASTUKSET

- Lisää perl-esimerkkejä:
  - prxparse("/\D+/") /* sisältää ei-numeroita */
  - prxparse("/^\D/") /* alkaa ei-numerolla */
  - prxparse("/^\d{6}[\A\+\-]\d{3}\w /") /* hetu: alkaa kuudella numerolla, sitten A,+ tai -, sitten 3 numeroa, lopuksi kirjain tai numero ja sitten tyhjää */
  - prxparse("/^\d{7}-\d /") /* y-tunnus: alkaa seitsemällä numerolla, sitten viiva ja numero ja tyhjää */
  - prxparse("/^\+358\d{6}/") /* alkaa +358 ja sitten ainakin 6 numeroa */
  - prxparse("/[^0-9^\-]/") /* sisältää muuta kuin numeroita tai alaviivoja */
  - Mahdollisuuksia valtavasti, lisää esimerkkejä esim:
    - ***http://www2.sas.com/proceedings/sugi29/265-29.pdf***

S-Pankki

CHAR RAKENNETARKASTUKSET

```
data work.analyysi ;
    set kaikki_yhteen;
    where tarkistustyyppi="Rakenne";
    p_ytun = prxparse("/^\d{7}-\d /");
    p_ei_voi_olla_ytun = prxparse("/[^0-9^\-]/");
    p_kotim_puhelin = prxparse("/^\+358\d{6}/");
    if lowcase(parametri1)="Y-tunnus" then do;
        if prxmatch(p_ytun,arvo) = 0
            or prxmatch(p_ei_voi_olla_ytun,arvo) > 0 then suspekti=1;
    end;
    else if lowcase(parametri)="Puhelin" then do;
        if prxmatch(p_kotim_puhelin,compress(arvo)) = 0 then suspekti=1;
    end;
run;
```

Kirjasto	Taulu	Muuttuja	Tarkistustyyppi	Parametri1	Arvo	Suspekti
Lib1	Taulu1	Var2	Rakenne	Puhelin	+358401234567	
Lib1	Taulu1	Var2	Rakenne	Puhelin	+35850ÖÖÖ	1
Lib2	Taulu3	Var9	Rakenne	Y-tunnus	1234567-8	
Lib2	Taulu3	Var9	Rakenne	Y-tunnus	123-45678	1

S-Pankki

CHAR ARVOJOUKKOTARKASTUKSET

- SASilla valtavasti tapoja verrata datan sisältöä toiseen tauluun:
  - hashit, formaatit, data step merge if not, sql arvo not in jne jne
- Oma esimerkkimme on melko yksinkertainen macro-looppaus sql:llä
  - Makrotetaan vertailua ohjaavat kentät
  - Luupataan kullekin, ja deletoidaan ne jotka löytyvät
    - Jäljelle jäävät vain suspektit
  - Yllättävänkin nopeaa
    - Osasyynä datan keräys yhteen kenttään, jolloin vältytään useilta samaan tauluun kohdistuvalta kyselyltä

S-Pankki

CHAR ARVOJOUKKOTARKASTUKSET

```
/* Haetaan tarkastuksen verrokkitaulu-kenttä -kombinaatiot */
proc sql;
    create table vertailukentat as select distinct vert_kirjasto, vert_taulu, vert_kentta from arvotarkistukset;
quit;
%let vertkentta_obs=&sqlobs.;
proc sql noprint;
    select vert_kirjasto, vert_taulu,vert_kentta into :vert_lib1 - :vert_lib&vertkentta_obs.,
                                                      :vert_taulu1 - :vert_taulu&vertkentta_obs.,
                                                      :vert_kentta1 - :vert_kentta&vertkentta_obs. from arvotarkistukset;
quit;

/* Luupataan tarkistukset läpi */
%do i=1 %to &vertkentta_obs.;
    proc sql;
        delete from arvotarkistukset
            where vert_kirjasto="&&vert_lib&i" and vert_taulu="&&vert_taulu&i" and vert_kentta="&&vert_kentta&i"
            and arvo in (select &&vert_kentta&i from &&vertkirjasto&i...&&vert_taulu&i
                where valid_from_dttm<=datetime()<=valid_to_dttm);
        quit;
%end;
```

012 ja 013 löytyvät tuotetaulusta, joten ne siivotaan pois, vain suspekti jää

Kirjasto	Taulu	Muuttuja	Tarkistustyyppi	Parametri1	Arvo
Lib2	Taulu3	Var8	Arvojoukko	Libx.tuotetaulu.tuotekoodi	ÖÖÖ

S-Pankki

LOPPUTULEMA

Kirjasto	Taulu	Muuttuja	Tunnusluku	Hälyraja	Arvo_eilen	Arvo_tänään	Pros_ero
Lib1	Taulu1	Aikaleima	Avg	<=0	14mar2016	14mar2016	0
Lib2	Taulu2	Euroja2	Max	>5	100 000	107 000	7

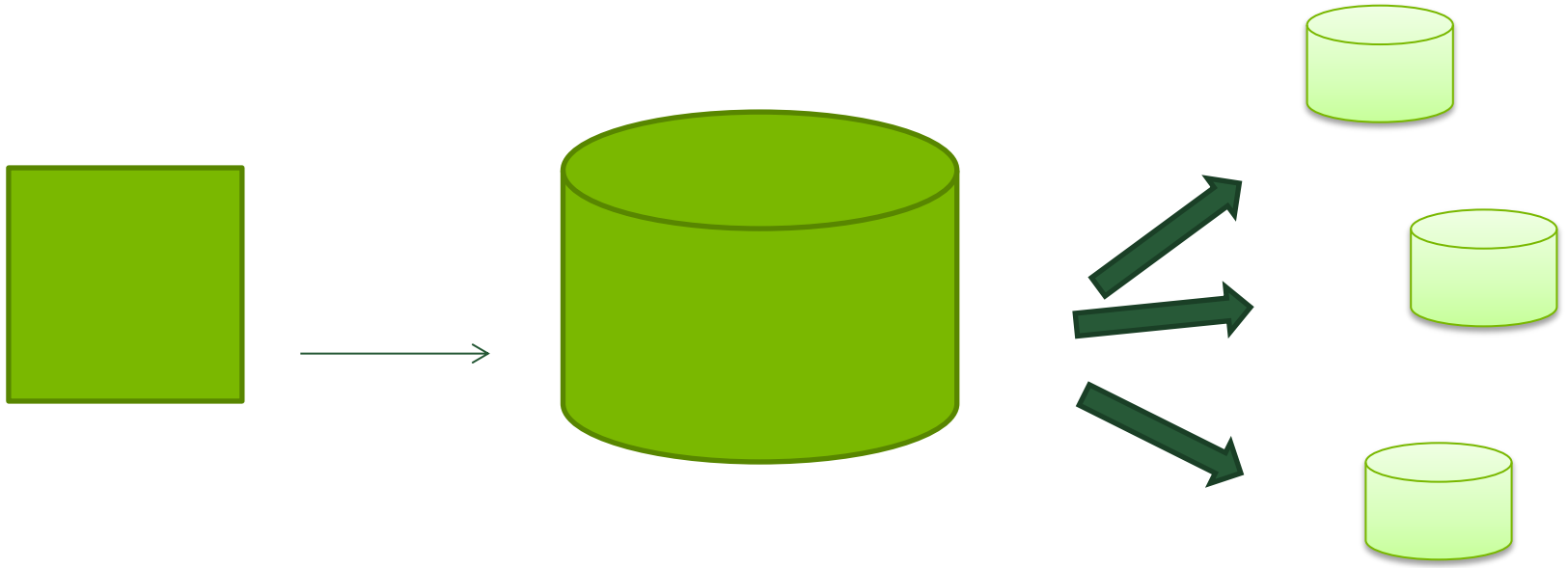
Kirjasto	Taulu	Muuttuja	Tarkistustyyppi	Parametri1	Arvo
Lib1	Taulu1	Var2	Rakenne	Puhelin	+35850ÖÖÖ
Lib2	Taulu3	Var9	Rakenne	Y-tunnus	123-45678

Kirjasto	Taulu	Muuttuja	Tarkistustyyppi	Parametri1	Arvo
Lib2	Taulu3	Var8	Arvojoukko	Libx.tuotetaulu.tuotekoodi	ÖÖÖ

- Jäljellä pelkät suspektit
  - Kaikista lib-, table-, var- ja arvotiedot, sekä suspektin käräyttäneen tarkistuksen tiedot
  - Helppo raportoida, lähettää eteenpäin, tehdä graafia yms

S-Pankki

TIEDON TARKISTAMINEN



tarkistus

S-Pankki

MARTTIEN TARKASTAMINEN

- Marttien poikkeusarvot kertovat ennen kaikkea käsittelyssä tapahtuneista virheistä
  - Väärin mäpätyt kentät
  - Virheet arkkitehtuurissa
- Marttien seuraaminen antaa selkeän kuvan muutoksista, joita ympäristössä tapahtuu ennen, kuin täsmäytetty tieto on käytettävissä.

S-Pankki

JALOSTETUN TIEDON TARKISTAMINEN

- Jos tarkistus tehdään DataMart tasolla, voidaan nähdä virheet, joita oma ETL prosessi on aikaansaanut.
- DataMartin tarkastus voi olla samanlaista ’virtatarkastelua’ kuin mitä on tietovaraston tarkastelu tai aikaa voidaan käyttää enemmänkin luokiteltujen asioiden tarkastamiseen.
- Marttien datan tarkastelu ei ole niin aikakriittistä, kuin datamassan lataamisen tarkastelu. Virhe martissa ei pysäytä kaiken taloon tulevan datan lataamista.

S-Pankki

IDEOINTIA KÄYTTÖMAHDOLLISUUKSISTA

- Miksi torstaisin tulee ’huonompia asiakkaita’ kuin maanantaisin?
  - Tiettyjen ’epäoleellisten’ arvojen keskittyminen joihinkin ajankohtiin voi indikoida asiakaskäyttäytymistä.
  - Minimien, maksimien ja keskiarvojen tarkkailulla päästään yllättäviin johtopäätöksiin.

S-Pankki

MITÄ EI OLE VIELÄ RATKAISTU

- Työnohjausjärjestelmään integroitu automaattinen prosessi olisi kiva..
- Kellotus ongelman esille tulosta kuittaukseen ja ratkaisuun asti.
- Tietopankki siitä, miten data oikein käyttäytyy ja millaisia ratkaisumalleja ongelmiin voidaan löytää.

S-Pankki