
kapsch
challenging limits
Kapsch CarrierCom
PRILAGODBA MODELA ZA SLOŽENE ANALIZE PODATAKA
Kapsch FMS 4.1

Tko smo mi?

- Kapsch CarrierCom
  - Vodeći, globalni proizvođač, dobavljač i sistem integrator *end-to-end* telekomunikacijskih rješenja.
  - Nudimo inovativne, poslovne proizvode, tehnologije i usluge za željeznički i javni gradski prijevoz, telekomunikacijske mreže te dobavljače energije.
  - Uz pomoć naših 9 R&D centara u Europi i Aziji, konstantno pomičemo granice tehnologija.

Automatic Fare collection

RC900 Cab Radio

Smart Tram Management

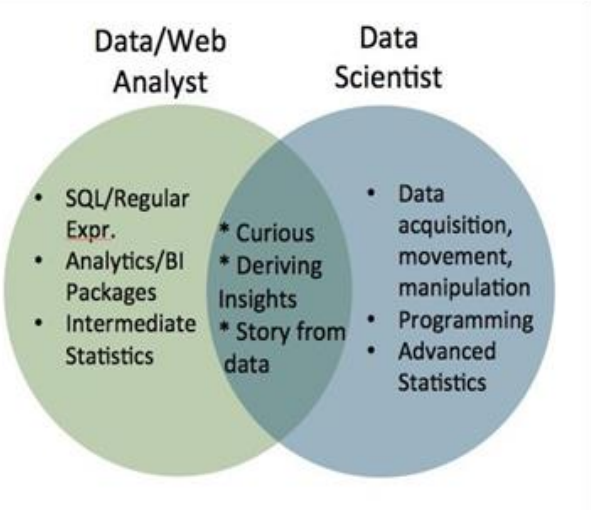
fraud management

Kratki uvod u Big Data

- BIG DATA analize
  - Primjenjuje se u svim relevantnim segmentima života i industrije

- DATA SCIENCE
  - Razvojem informatičkih tehnologija utrt je put još složenijim i širim analizama podataka

- DATA ANALYSIS vs. DATA SCIENCE
  - Razlika u načinu pristupa, tj. u znanjima i sposobnostima samih ljudi koji se bave jednom od te dvije djelatnosti



> BIG DATA

- Data Mart, Data Silos, itd…→ opisuju razne oblike kvalitetno, ali i loše pohranjenih, najčešće prethodno obrađenih podataka za koje se od analitičara podataka očekivalo da složenim transformacijama i analizama iznjedre nove, neočekivane korelacije i zaključke te značajno pridonesu uvećanju poslovanja i dobiti tvrtki, ili unaprjeđenju uvida i znanja.
- ➔ potrebni su PODATKOVNI ZNANSTVENICI

Vrste podataka i njihova pohrana

Sirovi

- Obradom dolazimo do informacija, ali se pohranjuju u sirovom obliku

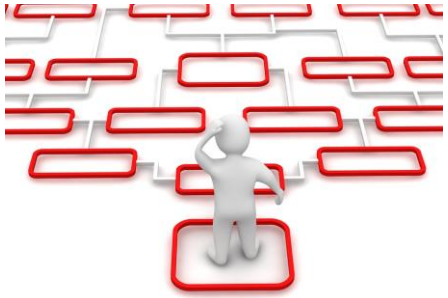
Nestrukturirani

- Nakon obrade i strukturiranja pohranjuju se u strukturiranom obliku

Kripto-strukturirani

- Sadrže u sebi strukturiranost, ali se do te strukture ne može doći bez strojne obrade

Strukturirani

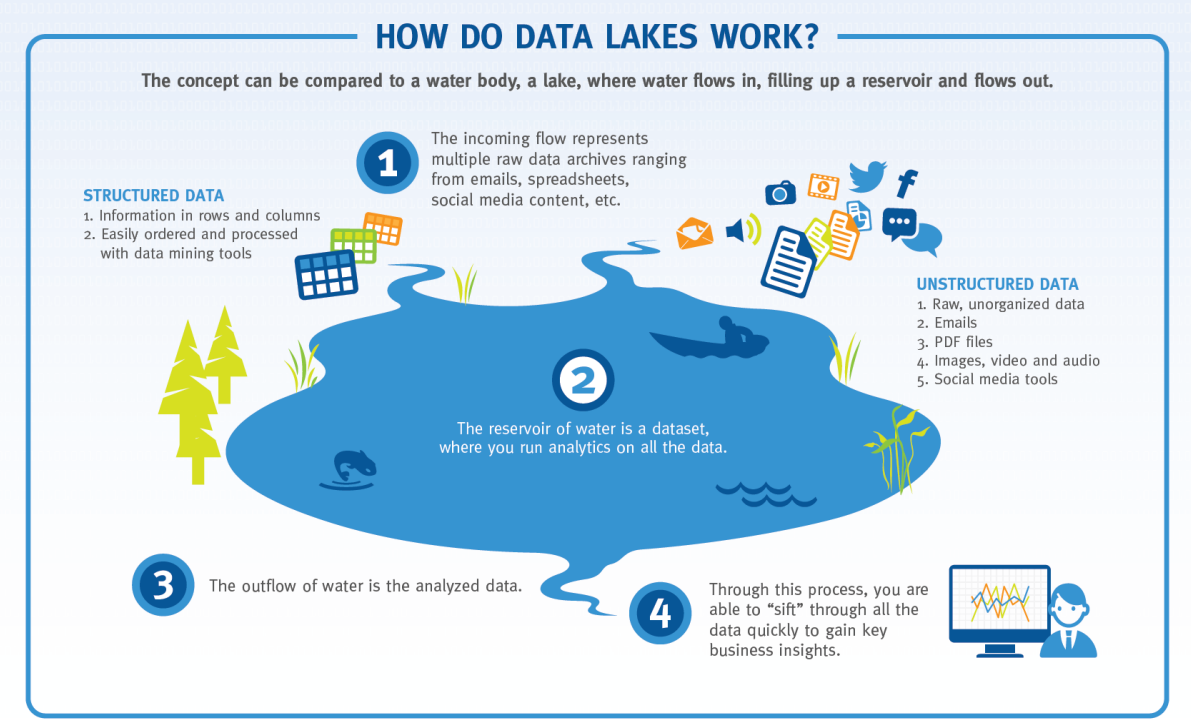


- Prikupljaju i spremaju u strukturiranom obliku

Ograničenja kod pohrane podataka

POHRANI SVE!!!

- DATA SILOS – loš primjer pohrane → lako za kreirati ➔ teško za održavati i koristiti
- **DATA LAKE** – ujedinjuje kapacitet i pohranu potrebnu za kvalitetnu analizu!!!



- NOVAC → HW komponente, licence…

FMS

Fraud Management System 4.1

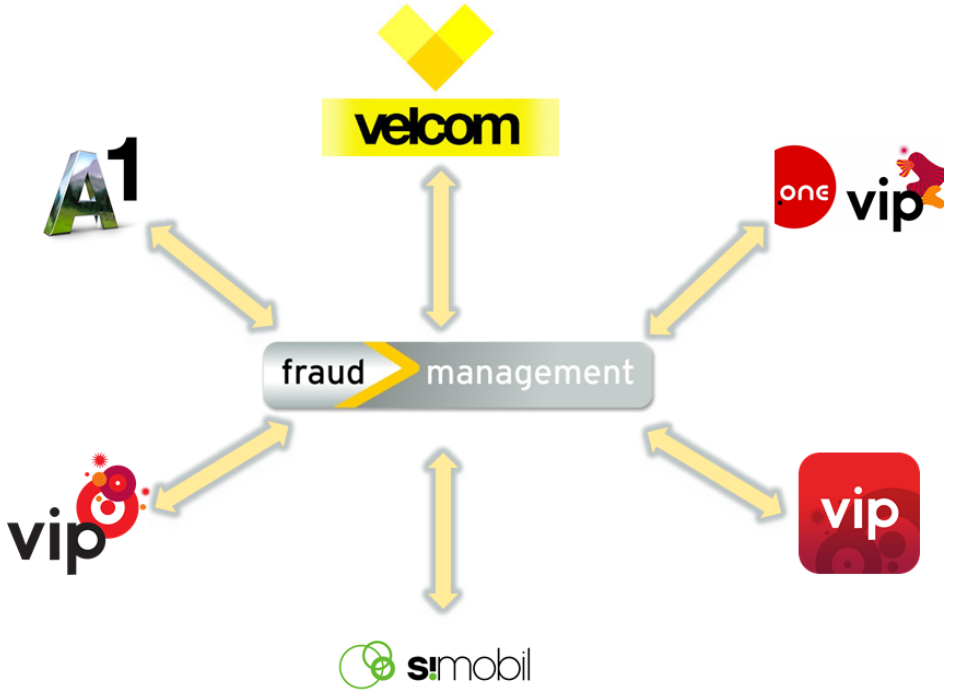
- Štiti telekomunikacijske operatere od prijevara (gubitka novca)

- FMS 4.1
  - 13 mil. korisnika
  - 226 mil. usluga
  - 170 mil. slogova/dan
  - zadržavanje podataka u RDBMS: 1-6 mjeseci

Zašto bi Vas mogao zanimati?

Fraud Management System 4.1

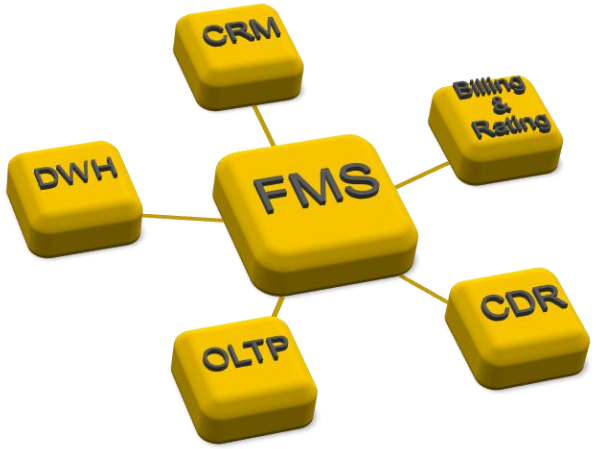
- Big Data
  - NEAR REAL TIME obrada
    - Obrada bez vremenske zadrške - ključna za verifikaciju
    - Kapsch FMS - dnevna obrada cca. 170 mil. slogova
- Primjer iz stvarnog svijeta
- Implementacija je rezultat
  - 15 godina iskustva na starom FMS-u
  - Rješavanje nakupljenih problema dobrom arhitekturom
  - Korištenje naprednih Oracle funkcionalnosti



Zašto bi Vas mogao zanimati?

Fraud Management System 4.1

- Podatkovni znanstvenici
  - Centralno mjesto prikupljanja podataka
    - Ogromna količina različitih podataka se slijeva u FMS
    - Data Lake
  - Relacijski model baze
    - Jednostavan za korištenje podataka
  - Pretprocesirani podaci
    - 80% vremena troše na pretprocesiranje



Zahtjevi za kvalitetnu znanstvenu analizu

Dostupnost podataka

Unificiranost i dosljednost

Dosljedno izdvajanje specijalnih slučajeva

Dehardkodizacija

Pročišćenost od grešaka

Fizička pohrana i logička povezanost

- potrebno je omogućiti najjednostavniji dohvat i najbržu obradu

Slijednost

Kvalitetna priprema okružja

- automatizacija repetitivnih procedura
- jednostavna i logična denormalizacija podataka - iznimno važno podatkovnim analitičarima i znanstvenicima
- obrade temeljene na podacima (data-driven obrade)

Modularnost i prilagodljivost

- model i automatizirane procedure prilagodljive budućim potrebama

Fizičko modeliranje pohrane

Matični podaci – podaci o korisnicima, šifrarnici i indeksi

Problemi

- Čuvaju se svi podaci
  - Deaktivirani
  - Aktivirani
- Koriste se u gotovo svim upitima
- Sporo pretraživanje od vrha prema dolie

Rješenja

- Logičko odvajanje particioniranjem češće korištenih od rjeđe korištenih podataka.
- Podaci i indeksi nužni u obradi stavljeni su u BUFFER_KEEP (32GB) – 3x ubrzanje učitavanja
- Izbjegavanje hijerarhije tipova

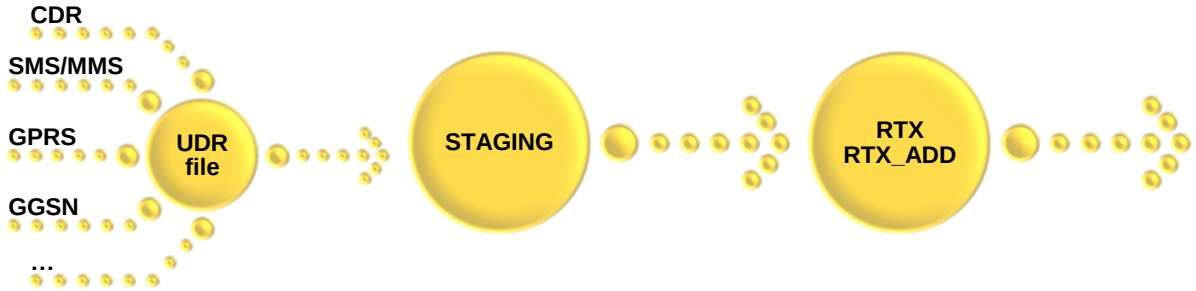
```
RESOURCES R JOIN DEF_IDENTIFIERS I
ON R.IDENT_ID = I.TIP
WHERE I.NADTIP = 'MSISDN'
```

Fizičko modeliranje pohrane

Prometni podaci

Problemi

- *Velika količina podataka za obradu u realnom vremenu*
- *Različiti tipovi podataka*
- *Različit obujam s obzirom na tip podatka*
- Potrebno raditi dvije vrste obrada:
  - obrade samo na *novo pridošlim podacima* te
  - *obrade za određeni vremenski period*



Rješenja

- Automatiziran i konfigurabilan sustav particioniranja
  - Vremensko
  - Logičko - s obzirom na tip podatka
- Korištenje STAGING tablica
  - Obrade brže nego na eksternim
  - Ujednačena u obliku i formatu za sve tipove
- Većina podataka završavaju u 2 tablice
  - Osnovni podaci iz datoteka (RTX)
  - Dodatni nastali izračunima (RTX_ADD)
  - INSERT /*+ APPEND */ SUBPARTITION FOR () – lock subpart.

Fizičko modeliranje pohrane

Agregirani podaci

Problemi

- Agregiranje po:
  - Sat
  - Dan
  - Mjesec
- Problem čuvanja podataka i njihovog dohvaćanja
  - Mjesečne particije znaju poprilično narasti
- Broj usluga stalno raste pa je potrebno napraviti dinamičko subparticioniranje

Rješenja

- Korištene IOT tablice
  - Brži dohvat (bez INDEX by ROWID)
  - Oracle ne podržava subparticioniranja IOT tablica
- Razvijen vlastiti mehanizam subparticioniranja pomoću Oracle multi-column particioniranja
  - Subparticioniranje po IDu korisničke usluge
  - Procedure za koje prate rast IDeva usluga

PARTITION BY RANGE (DATE_DAY, SERVICE_ID)

Dostupnost podataka

Problemi

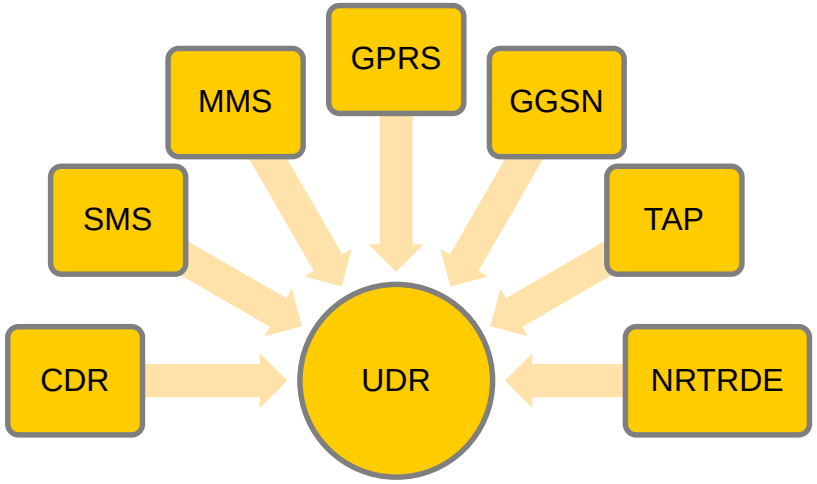
- Za kvalitetnu znanstvenu analizu potrebni su SVI podaci
  - Problem količine podataka i njihove pohrane

Rješenja

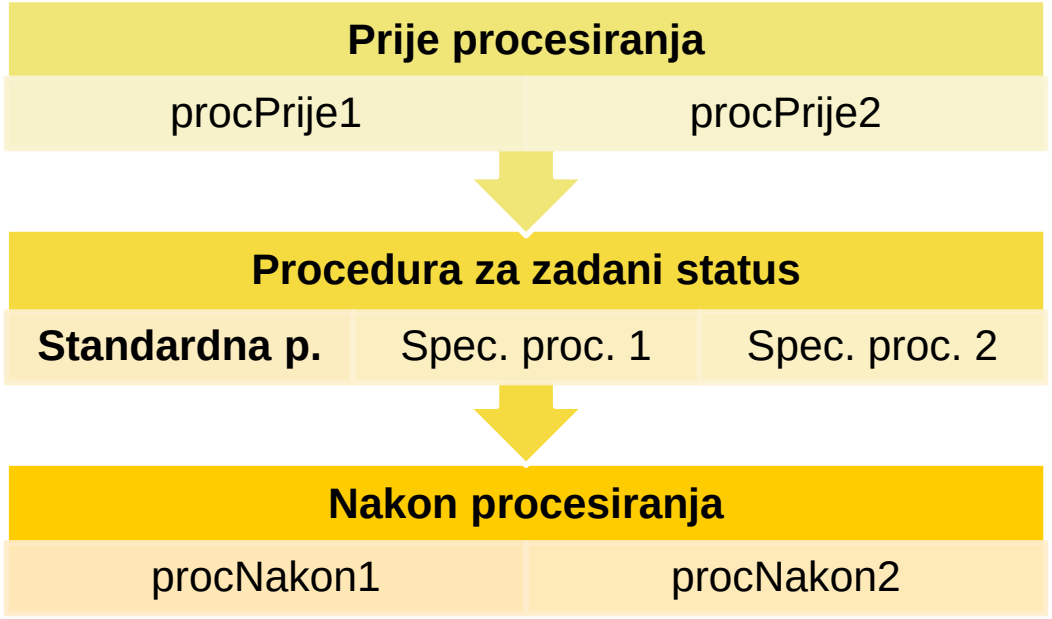
- Pohrana komprimiranih, često binarnih datoteka umjesto slogova podataka
- Upotreba SCHEMA_ON_READ arhitekture – dohvat sirovih podataka na osnovu sheme zadane u trenutku dohvata
  - Trenutačno se radi na doradi Kapsch FMS-a kako bi podržao i tu mogućnost što će rezultirati:
    - Smanjivanjem troškova pohrane
    - Mogućnošću dohvata podataka izvan definiranog konteksta → omogućen rad kako *Fraud* agentima tako i podatkovnim znanstvenicima.

Dorade modela podataka

- Unificiranost
  - Fizička
  - Logička



- Specijalni slučajevi konfigurabilni s obzirom na:
  - Tip podatka
  - Status



Dorade modela podataka

- Jedinstvenost i izdvajanje podataka
  - Duplikati
  - Nepotrebni slogovi

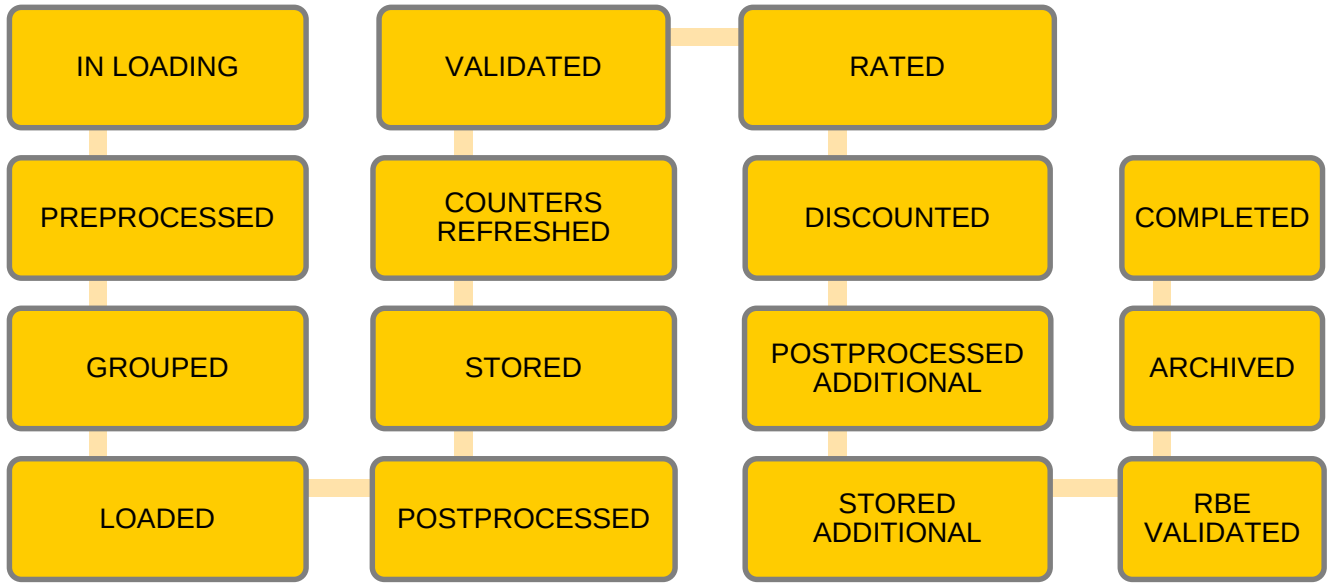
- Verifikacija (čišćenje podataka)
  - Pogrešni ulazni podaci
  - Greška unutar sustava

- Dehardkodizacija
  - Pozivi algoritama za specijalne slučajeve ➔ obrade temeljene na podacima
  - IF-THEN-ELSE blokovi ➔ kreiranje šifrarnika
  - Kompleksno mapiranje ➔ post procesirajućom procedurom uvesti tip u sustav bez mijenjanja standardne procedure
  - Izračun šifre ➔ zapisati u bazi

IF izračunajTip (šifra_usluge, b_broj, grupa) THEN

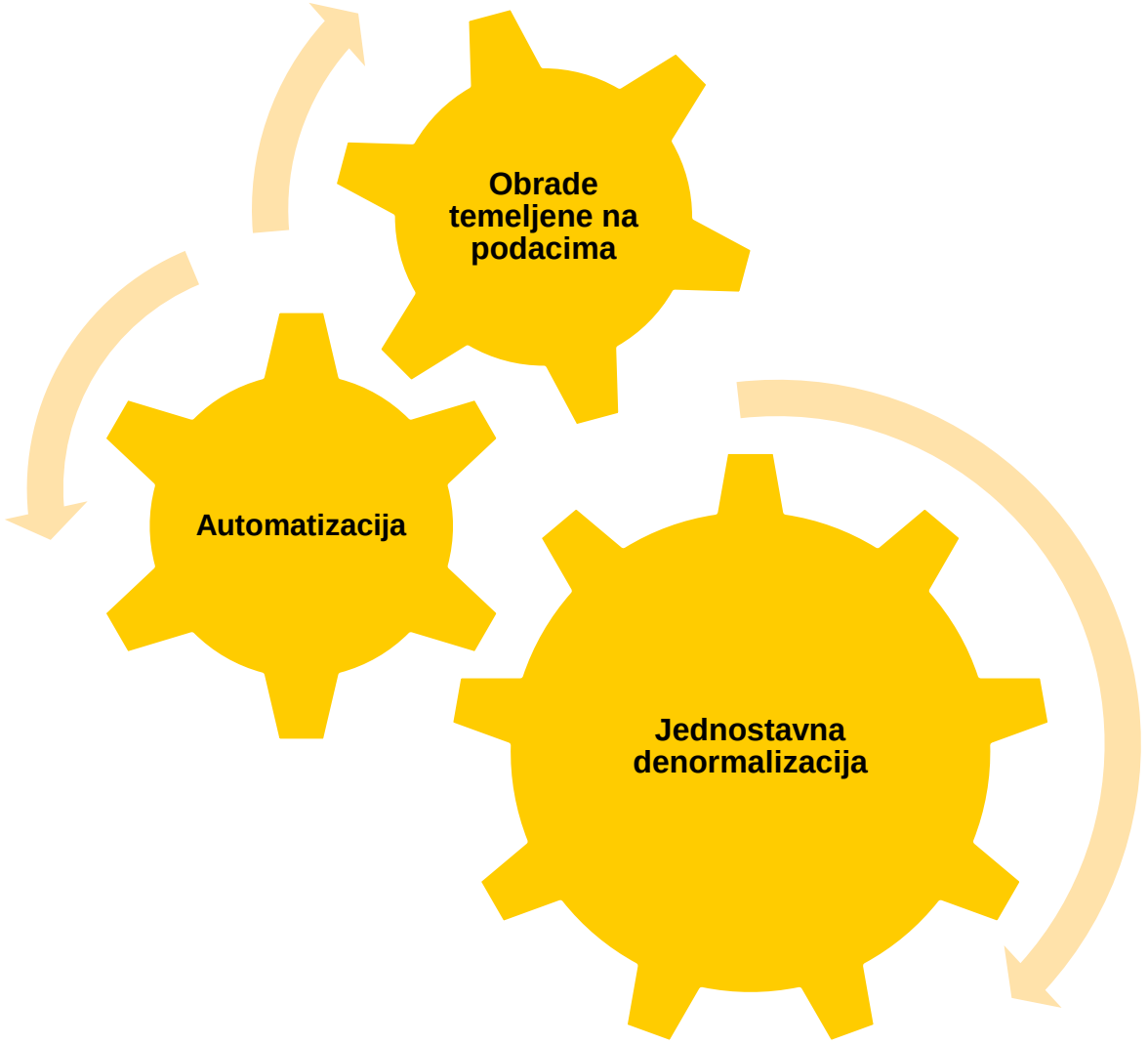
Dorade modela podataka

- Slijednost podataka
  - Slijeđenje svake transformacije, agregacije, događaja ili akcije do njenog samog uzročnika
  - Historijat šifrarnika i matičnih podataka
    - Jedna tablica (*dictionary nested* tablica, XML polje)
    - Uzročnik promjene
    - Vrijeme promjene
  - STAGING tablice – podatak o preduzročniku
  - Agregacije
    - performansna i prostorna ograničenja
    - Datum zadnje izmjene brojača
  - Vanjski sirovi podaci (datoteka, WS, …)
    - Dostupnost
    - Statistička tablica – LOAD_STATISTICS
      - Vremena pojedinih dijelova procesiranja



Dorade modela podataka

- Kvalitetna priprema okružja
  - Uvođenje šifrarnika
  - Uočiti učestale zadatke
  - Koristiti pakete u bazi
  - WORK tablespace
    - Privremeni podaci
    - DBA ne mora raditi backup



Zaključak

- Mnoga od predloženih rješenja ne zvuče kao nešto novo
  - Kombinacijom postojećih poznatih rješenja sa novonastalim potrebama čine potrebnu dodanu vrijednost
- Neka od rješenja mogu djelovati kao pretjerivanje (samoverifikacija sustava)
  - Praksa pokazuje upravo suprotno
- Neka rješenja su možda preambiciozna

- ➔ robusni sustav koji je lako za održavati i uvodi Kapsch FMS na velika vrata u BIG DATA svijet

Reference

- https://en.wikipedia.org/wiki/Big_data
- https://en.wikipedia.org/wiki/Data_analysis
- https://en.wikipedia.org/wiki/Data_lake
- https://en.wikipedia.org/wiki/NewSQL
- https://www.techopedia.com/definition/30153/schema-on-read
- http://www.business2community.com/big-data/top-5-problems-big-data-solve-01597918
- https://www.import.io/post/data-scientists-vs-data-analysts-why-the-distinction-matters/
- https://www.digitalnewsasia.com/business/forget-data-warehousing-its-data-lakes-now
- http://insights.dice.com/2014/03/05/data-science-is-dead/
- https://www.wired.com/2013/09/nsa-backdoor/
- http://www.orchardsoft.com/orchard-pathology/structured-data-for-synoptic-reporting/
- http://www.conductor.com/blog/2013/01/why-structured-data-should-be-in-your-2013-seo-strategy/
- http://blog.liquidhub.com/2015/06/data-science-enabled-customer-service-with-service-cloud-new-intelligence-engine
- https://www.mhbank.com/webres/Image/Learning/Fraud%20Protection.jpg
- http://www.lekarnaljubljana.si/public/datoteke/cebelice2.png
- https://www.backupassist.com/blog/wp-content/uploads/2013/04/iStock_000000393598_L3-700x363.jpg
- http://www.techsophy.com/wp-content/uploads/2015/01/041991316767-1.png
- http://19gdvc3ur026r9ray1n26t2q93.wpengine.netdna-cdn.com/wp-content/uploads/2015/04/overview-image1.png
- https://thumb7.shutterstock.com/display_pic_with_logo/643435/215663719/stock-vector-businessman-loss-money-during-carrying-because-of-careless-215663719.jpg

Autori

Hrvoje Devčić
Radnička 39 | 10000 Zagreb | Hrvatska
Phone +385-91-235-5820 | Fax + 385-1-457-3773
E-mail hrvoje.devcic@kapsch.net
www.kapsch.net/hr/kcc

Dario Nikolić
Radnička 39 | 10000 Zagreb | Hrvatska
Phone +385 91 896 00 63 | Fax + 385-1-457-3773
E-mail dario.nikolic@kapsch.net
www.kapsch.net/hr/kcc

Hvala na pažnji

Kapsch CarrierCom d.o.o.

Radnička cesta 39
10000 Zagreb, Hrvatska
Phone: +385 1 6408 838
Fax + 385-1-457-3773
www.kapschcarrier.com
www.kapsch.net