Pentaho Data Integration

5 miliona zapisa u 10 minuta, svaki dan

Goran Cvijanović

Vinteh d.o.o.

Sadržaj prezentacije

- Zablude oko implementacije skladišta podataka
- Arhitektura integracije podataka i ETL procesa
- Open source alati za integraciju podataka
- Pentaho Open source BI
- Primjena Pentaho Data Integration platforme
- Zaključak

Zablude oko implementacije SP-a

- Sustav koji prikuplja, pročišćava, verificira i isporučuje izvorišne podatke u dimenzionalna skladišta podataka, te osigurava i implementira upitne i analitičke module za podršku poslovnom odlučivanju
- Korisnika smještamo u centar razvoja skladišta podataka, značaj pojedinih komponenti ima drugačiji pogled
- Put do točnih i upotrebljivih podataka, ETL procesi
- Podsjetnik prilikom planiranja izgradnje skladišta podataka

Što SP zapravo nije

- Možemo li kupiti gotovo rješenje koje će biti naše skladište podataka?
  - Ne. Moramo ga izgraditi ili prilagoditi, jer naše poslovanje nije identično nekom drugom
  - Sam proces izgradnje uključuje analizu poslovnih procesa i izvorišta podataka, manipulaciju podacima, modeliranje dimenzijskih struktura i korisničkih sučelja
- Za izgradnju skladišta podataka ne koristi se jedinstven programski jezik
  - Njega čini više komponenti koje uključuju različite programske arhitekture
- Projekt skladišta podataka čine grupe manjih projekata i faza
  - Implementacija se izvodi u više nezavisnih projekata koji omogućavaju da se izgradnja osnovnog skupa funkcionalnosti provede do kraja
  - Time se omogućava da projekt skladišta podataka bude uspješan, upravljiv i izvediv u zadanom vremenu
- Model podataka ne predstavlja skladište podataka
  - Čak i najbolje osmišljen model je beskoristan bez kvalitetnih podataka i razumljive prezentacije
- Kopiranje produkcijskog sustava na novi poslužitelj zbog ubrzanja izvještavanja ne čini taj sustav skladištem podataka
  - Bez promjene strukture onemogućena je osnovna namjena skladišta podataka, podrške poslovnom odlučivanju

Arhitektura IP i ETL procesa

- Pogled od 360 stupnjeva na poslovne podatke, predstavlja moderni termin za pojam integracije podataka
- Postići zadovoljavajuću kvalitetu integriranih podataka, kroz osiguravanje kvalitetnog izvora transakcijskih podataka
- Usklada dimenzija s ciljem postizanja jednakih vrijednosti
- Usklada vrijednosti za mjere značajna je za mogućnost izgradnje takozvanih ključnih indikatora
- Izbor arhitekture je elementaran i potrebno ga je odraditi u ranoj fazi planiranja ETL sustava
- Dva osnovna puta, korištenje gotovog rješenja tehnološkog partnera ili izgradnja vlastitog ETL sustava

Prednosti i nedostaci pojedinog modela

GOTOVO RJEŠENJE

- Jednostavniji, brži i jeftiniji razvoj
- Efikasno korištenje alata i bez poznavanja o programskih jezika i programiranja
- Postojanje među-podatkovnog sloja (metadata) za postizanje uniformnosti
- Postojanje predefiniranih konektora za mnoštvo izvorišnih i odredišnih sustava
- Korištenje ugrađenih funkcija za sigurne komunikacije, enkripcije i kompresije
- Dobre performanse i podrška za rad s velikim količinama podataka
- Podrška za rad u klasteriranim i sustavima, osiguranje integriteta

Prednosti i nedostaci pojedinog modela

VLASTITI RAZVOJ

- Mogućnost integralnog procesa testiranja korištenjem alata za testiranje
- Objektno orijentirane tehnike omogućuju koegzistentnost i nadzor nad graškama
- Nezavisnost od proizvođača alata za ETL, kao i puna fleksibilnost u korištenju poznatih razvojnih alata i jezika

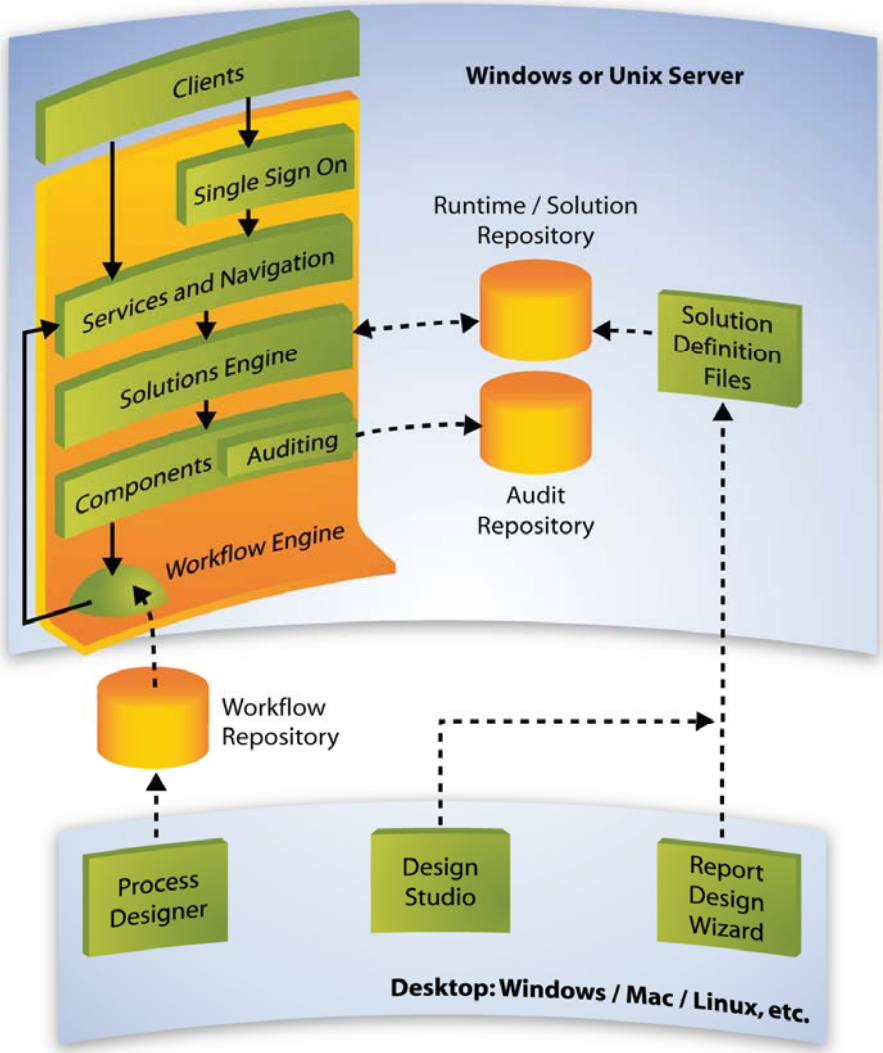
---

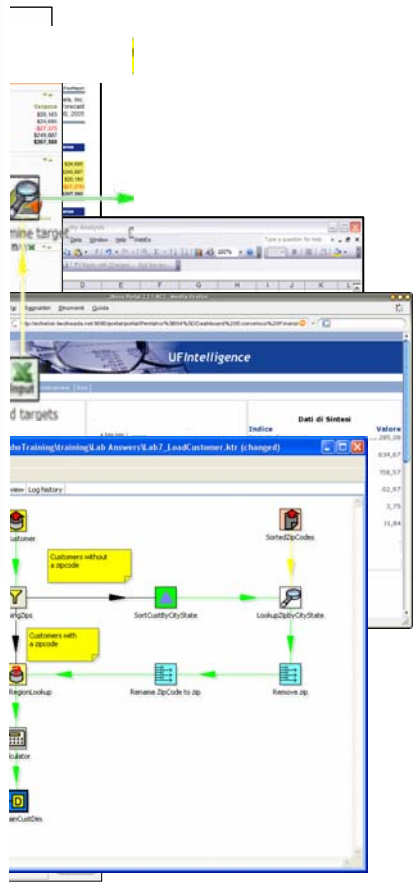
- Idealno bi bilo imati rješenje koje nudi osobine iz obije grupe, gotovo rješenje sa svim svojim prednostima i mogućnostima, ali i dostupnost izvornog koda uz poznavanje programske platforme na kojoj je izrađeno
- Kao takva mogućnost su open source rješenja koja predstavljaju već zrele produkte i rješenja koja mogu zadovoljiti široku paletu namjene

Open source alati i platforme za BI

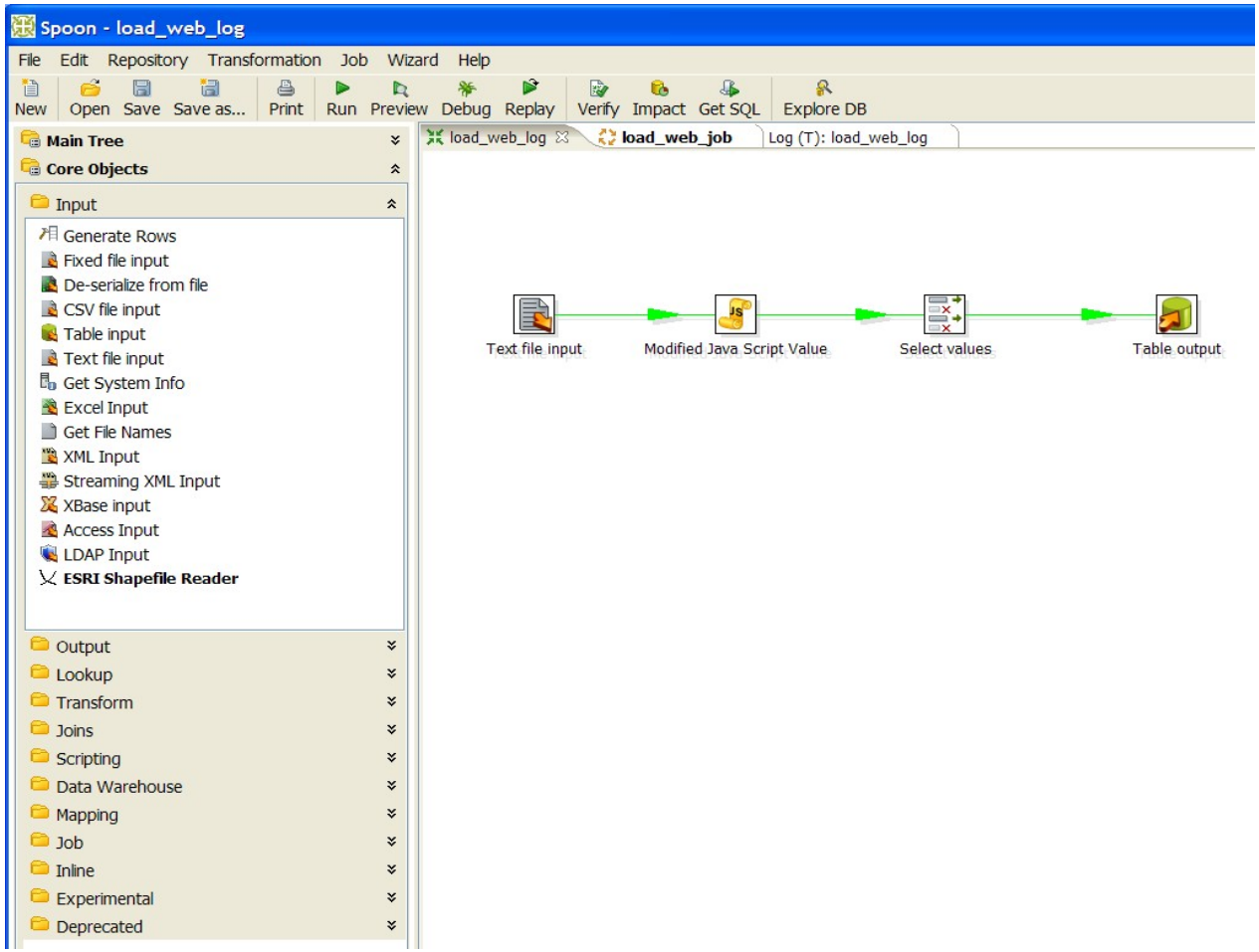
- Pentaho - najpoznatije open source rješenje, uključuje izvještajni sustav, analize, prezentacijsku platformu (dashboard) i rudarenje podataka (data mining).
- JasperSoft - poznato ime iz svijeta izvještajnih open source platformi, nastalo udruživanjem Jasper Reports i iReport platformi
- Actuate Corporation - BIRT platforma koja je dio Eclipse fondacije. Predstavlja platformu koja se intezivno razvija uz veliku podršku matične kompanije
- Spago BI - tvrtka koja nudi kompletno open source rješenje, u kombinaciji s profesionalnim implementatorima koji mogu izvesti složene projekte

Pentaho Open source BI





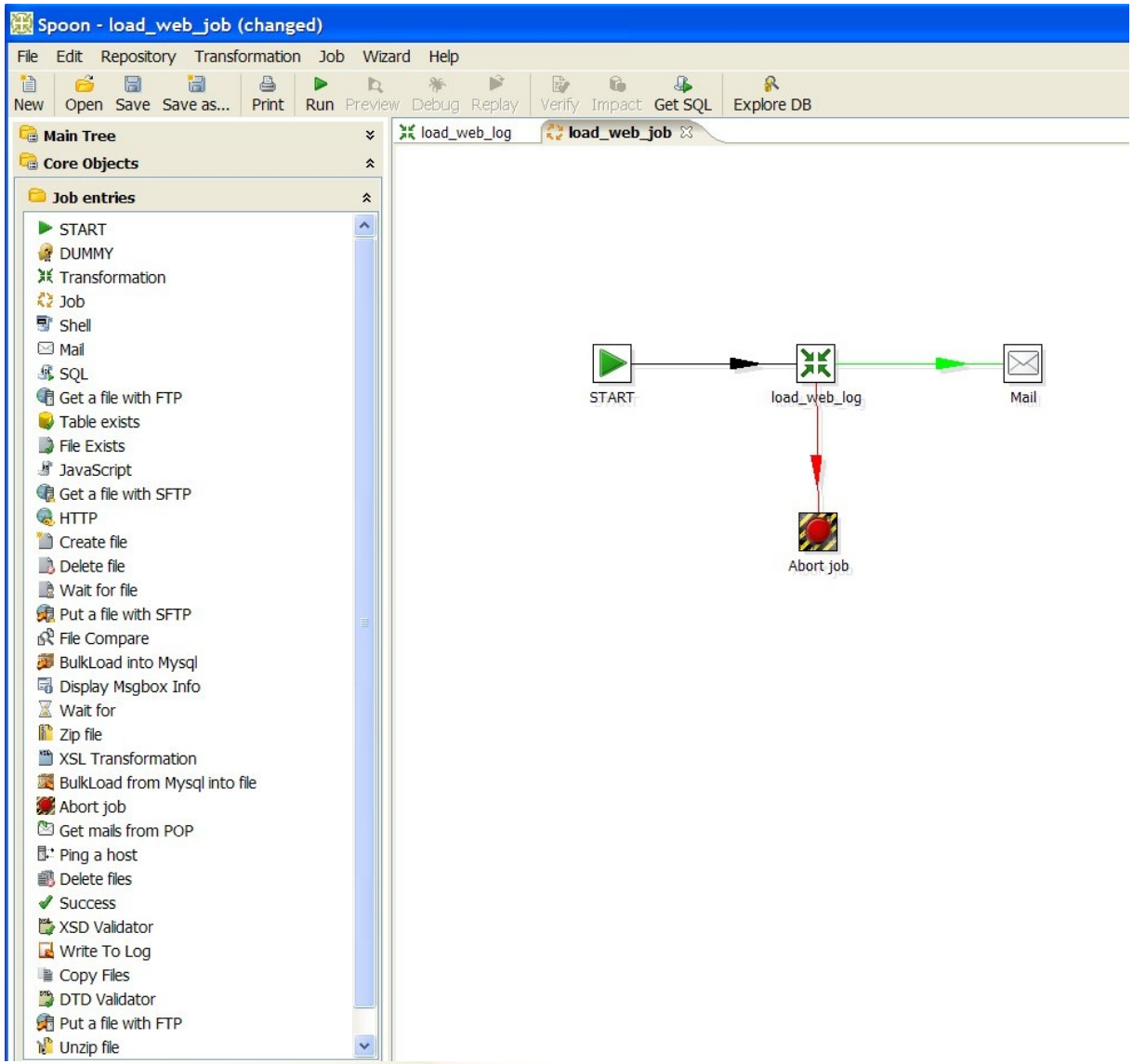
Primjena Pentaho Data Integration platforme



Upotreba

- Kao sistemska platforma koriste se Linux serveri
- Java Runtime Environment verzije 1.4 ili noviji
- Podržane platforme su: Microsoft Windows uključujući i Vista verziju, Linux, Apple OSx, Solaris, AIX, HP-UX, FreeBSD
- Spoon, Pan i Kitchen moduli aplikacijske platforme

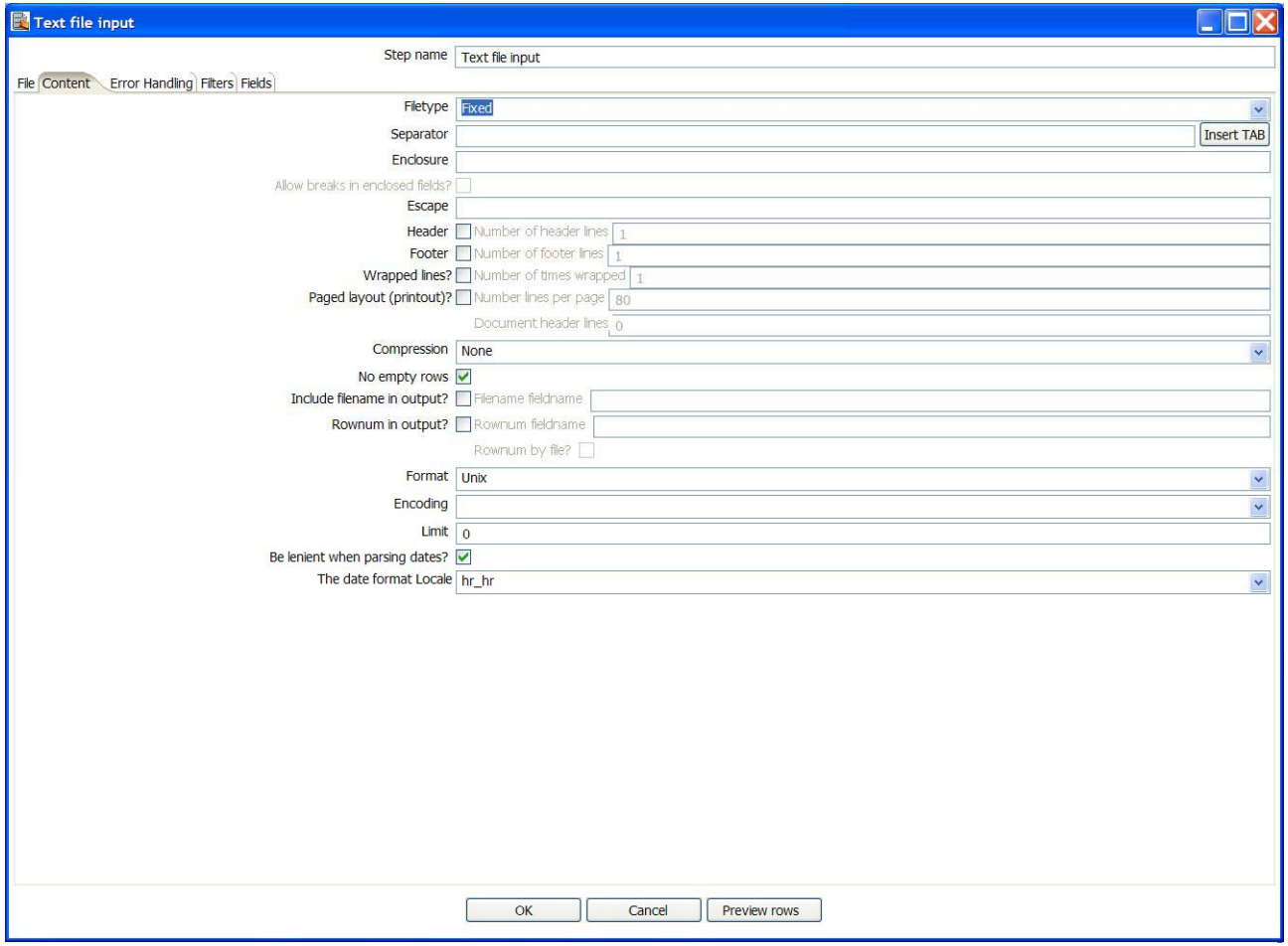
Prikaz PDI Spoon modula



Spoon modul

- Osnovni modul koji se koristi za modeliranje i izvršavanje transformacija i upravljačkih procesa
- Proces koji opisujemo je učitavanje web logova koji su standardne Apache Web Server strukture
- Izvršni proces sadrži elemente i njihove tokove, u ovom slučaju oznaku početka procesa, pokretanje transformacije, te grananje u slučaju greške prekid procesa, a u slučaju uspješnog izvršenja slanje mail poruke o obavljenom poslu

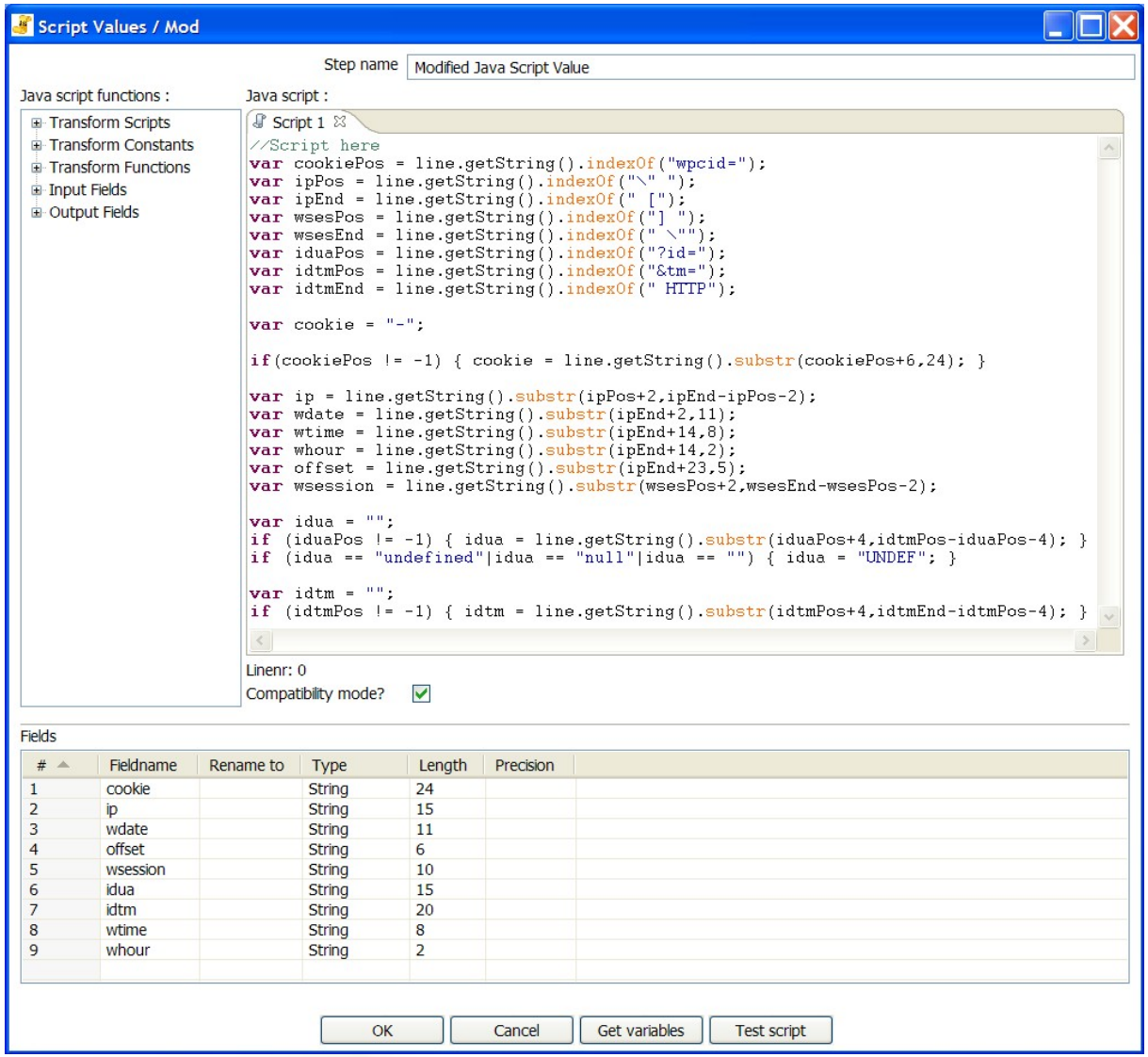
Izvedbeni koraci – učitavanje iz tekstualne datoteke



Opis

- Odabran je rad s redovima u datoteci koji su fiksne veličine, nisu dozvoljeni prazni redci, te se radi o Unix formatu završetka retka
- Učitava se cijeli jedan redak u jedno tekstualno polje, a u sljedećem koraku će se odraditi izdvajanje pojedinih elemenata iz teksta

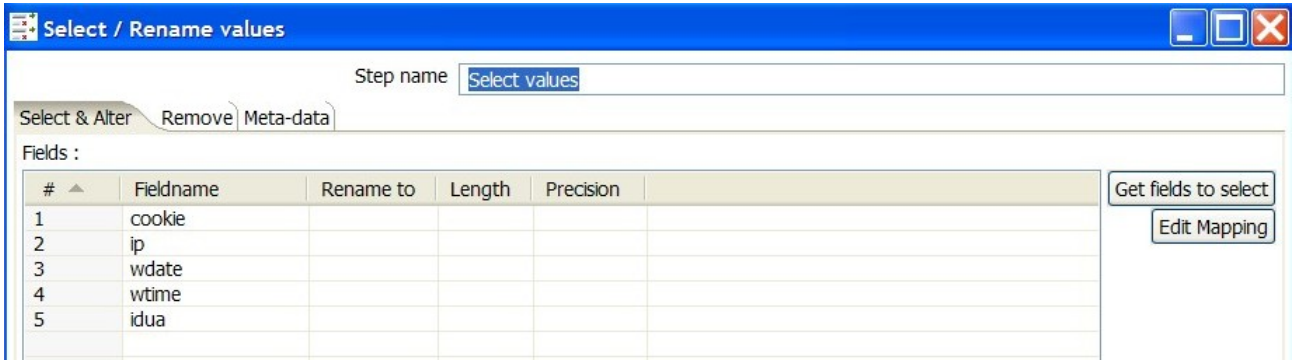
Izvedbeni koraci – Java script

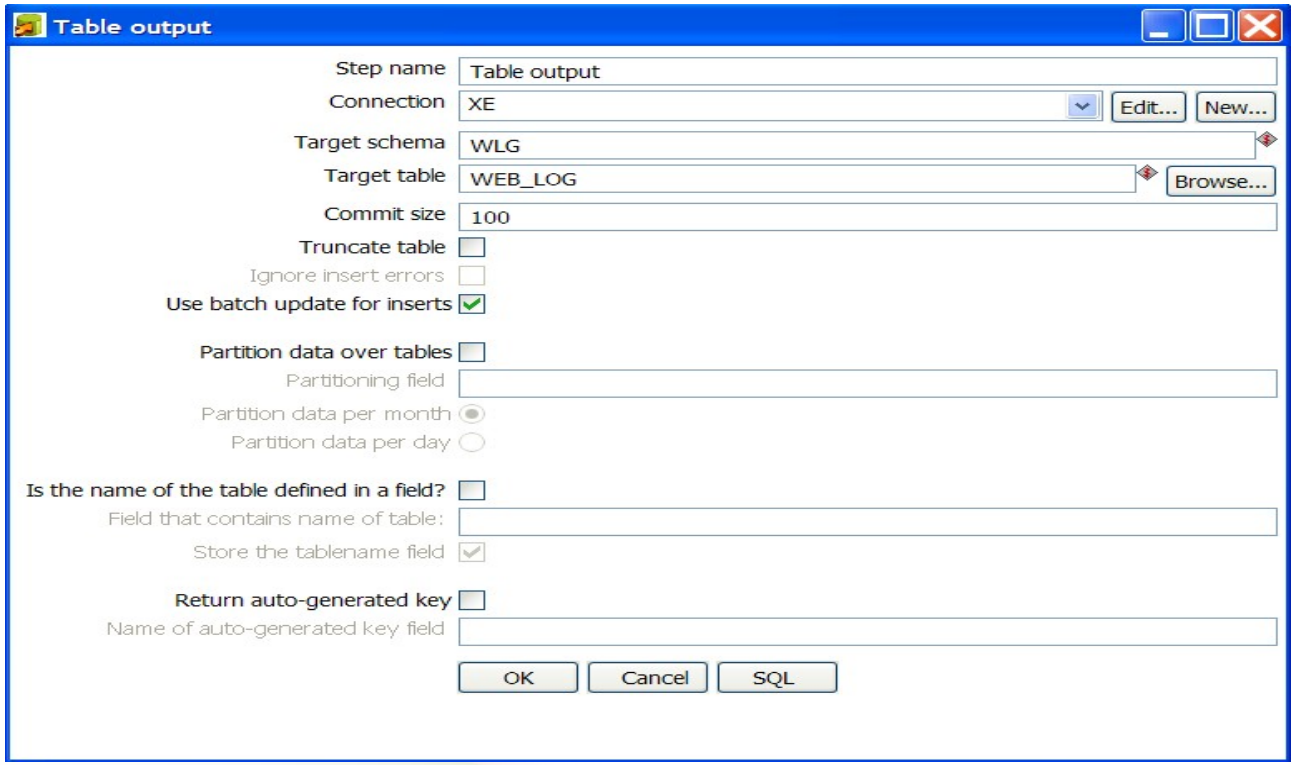


Opis

- Izdvajanje pojedinih elemenata iz teksta napravljeno je pomoću Java script jezika
- Elementarnom poznavanje sintakse i načina kako rastaviti tekstualno polje na pojedine elemente pomoću predefiniranih razdjelnika u sadržaju teksta
- Sva su polja usklađena sa strukturom kolona u tablici Oracle baze podataka

Izvedbeni koraci – mapiranje i pohranjivanje u bazu





Opis

- Mapiranje u oblik pogodan za spremane u bazu podataka
- S obzirom da smo vodili računa o veličini pojedinih varijabli i njihovom tipu, mapiranje se trivijalno svodi na popis varijabli koje imaju jednak naziv kao i kolone u tablici web_log Oracle baze
- Povezivanje s velikim brojem relacijskih baza podataka omogućeno je već nakon instalacije

Zaključak

load_web_log | load_web_job | Log (T): load_web_log

#	Stepname	Copynr	Read	Written	Input	Output	Updated	Rejected	Errors	Active	Time	Speed (r/s)	input/output
1	Text file input	0	0	4459547	4459547	0	0	0	0	Finished	398.4	11193.0	-
2	Modified Java Script Value	0	4459547	4459547	0	0	0	0	0	Finished	399.2	11171.9	-
3	Select values	0	4459547	4459547	0	0	0	0	0	Finished	399.2	11171.5	-
4	Table output	0	4459547	4459547	0	4459547	0	0	0	Finished	399.3	11167.1	-

Za napomenuti

- Bitni segmenti u fazama razvoja produkta su svježina ideja i smjernice razvoja, veliki broj korisnika koji aktivno testiraju i prijavljuju uočene probleme, podrška u rješavanju problema
- Pentaho Data Integration je alat koji omogućava da se poslovi koji čine ETL proces obave brzo i pomoću alata u kojem je ugrađeno znanje o tome kako nešto napraviti, a na stručnjaku je da odredi što treba napraviti
- kompletna podrška za zapisivanje informacija o odvijanju procesa, statističkih podataka interaktivno kroz alat, kao i mogućnosti pohranjivanja tih podataka u datoteke ili baze podataka koji se nadalje mogu analizirati ili prosljeđivati putem mail sustava

Pitanja za kraj