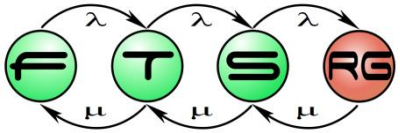
Vizuális adatelemzés - Gyakorlat



Adatelemzés szerepe a rendszermodellezésben

- Lényeges paraméterek meghatározása
  - Pl. mérési adatokból mik fontosak

  → szimuláció alapja
- Mérési adatok értelmezése
  - Milyen eredményt kaptunk valójában?
- Rendszerek összehasonlítása
  - Adott feladatra melyik alkalmasabb

A gyakorlat célja

- Hogyan találhatunk meg magas szintű összefüggéseket egy többdimenziós (mérési) adathalmazban?
  - „Egy adott frekvencia alapján lehet legjobban meghatározni az elhaladó jármű típusát”
  - „Bizonyos típusú algoritmusok jobban teljesítenek nagy modelleken, ha sok a változtatás”
  - „A kis konfigurációknál számít erősebben az OS”
- Ízelítő két gyakorlati problémából
  - Jó-e az általunk fejlesztett algoritmus?
  - Milyen adatbázis konfigurációt válasszunk?

Mit láttunk az előadáson?

- Diagram típusokat:
  - Oszlop
  - Pont-pont (scatterplot)
  - Scatterplot mátrix
  - Mozaik
  - Hisztrogram
  - Doboz
  - Párhuzamos koordináták
- Közelítő módszereket:
  - Regresszió, lineáris regresszió, simítógörbe
    - Láttuk a matematikai módszert is

Mit fogunk most látni?

- Végignézzük ezeket a diagramtípusokat egy valódi adathalmazon (benchmark eredmények)
- Látni fogunk példát a regressziós módszerek használatára

Milyen eszközök állnak rendelkezésünkre?

- R nyelv és a hozzákapcsolódó fejlesztőkörnyezetek
  - Bővíthetőség
  - Kiterjedt alkalmazási/statisztikai lehetőségek
- Mondrian adatvizualizációs szoftver
  - Képes RData állományok betöltésére
  - Képes R környezethez kapcsolódni (pl. regresszió)
  - Mi most (elsősorban) ezt fogjuk használni

MODELLKEZELŐ ALGORITMUSOK/ESZKÖZÖK ÖSSZEHASONLÍTÁSA

A feladat: modellek kezelése

- Modell: fogalmak és kapcsolatok (gráf)
- Kezelés: írás/olvasási műveletek
- Cél:
  - Tanszéki fejlesztésű eszköz (EMF IncQuery) és ipari eszközök hatékonyságának elemzése
  - Mintaillesztés és illesztések karbantartása
  - „Inkrementális”: ~ cache használata
- Esettanulmány (konkrét modell)
  - Vasúti irányítás
  - Modellméret: 1-10-100-1000 szorzók

Elemzési kérdések

- Szintetikus terhelés:
  - Egyszeri, batch jellegű módosítás (pl. validáció/kiegészítés transzformációval)
  - Sorozatos felhasználói módosítások
  - ~mintaillesztési benchmark
- Eszközök teljesítménye
  - Modellméret/feladat függvényében
  - Mit tudnak végrehajtani 10 mp alatt
- Eszközök erőforrásigénye

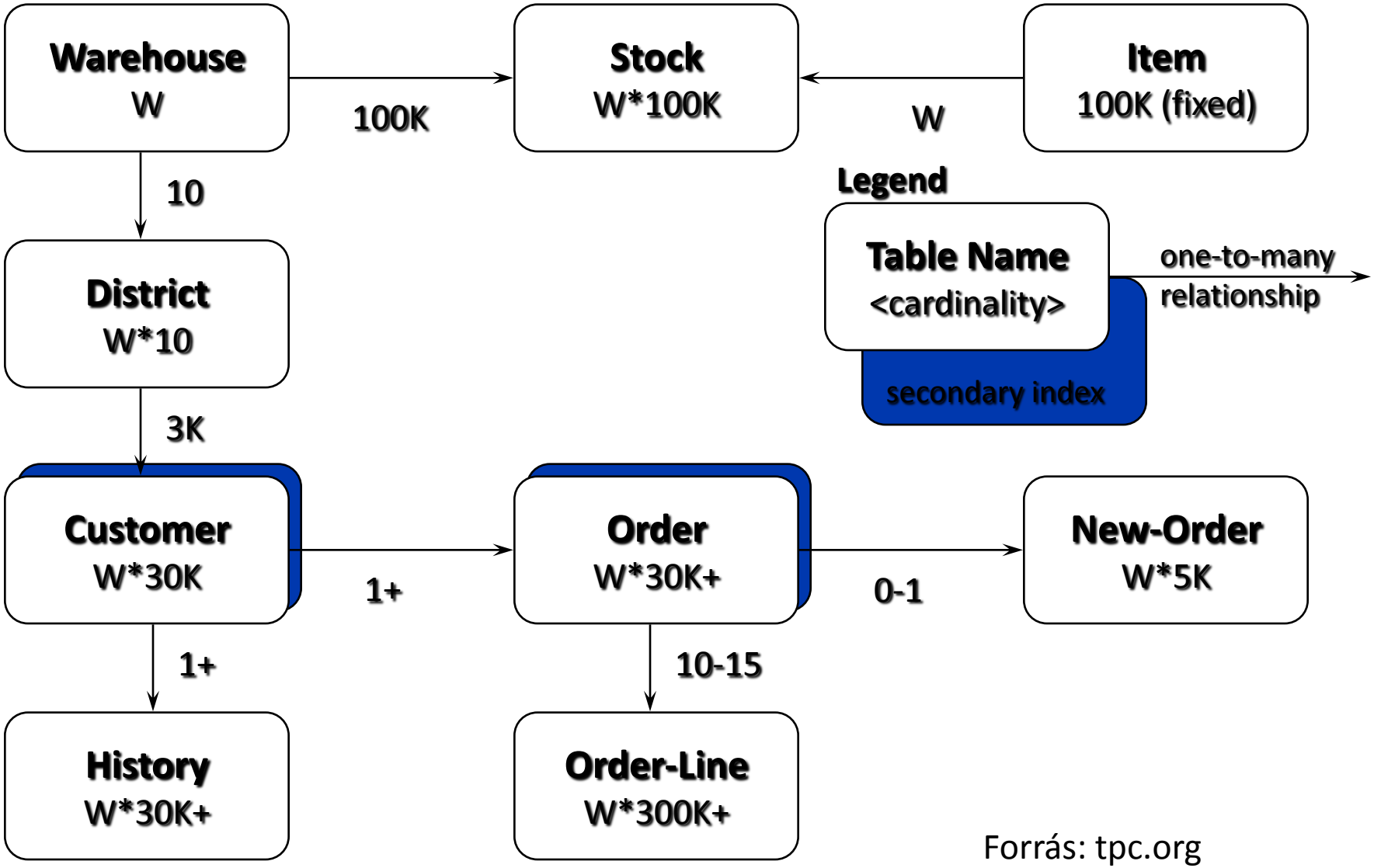
BEMUTATÓ (MONDRIAN)

BENCHMARK EREDMÉNYEK ELEMZÉSE

A TPC benchmark

- Adatbáziskezelő rendszerek mérése
  - RDBMS+OS+HW
- Mérési környezet
  - Mintaadatbázis: Ügyfelek és megrendelések
  - 5 fajta tranzakció (lekérdezés/módosítás) vegyesen
  - Felső korlát a futási időre
  - Valós körülmények: ACID tranzakciók, felhasználói gondolkodási idők
- Mért adatok
  - Áteresztőképesség (tpmC)
  - „Hatékonyság” ($/tpmC)

TPC-C séma



Forrás: tpc.org

A FELADAT

ADATTISZTÍTÁS

Cél: adathalmaz (tpc.org) előkészítése az elemzéshez

Adatok tisztítása manuálisan

- Kisebb méretű adathalmaz kézzel is tisztítható
- Előforduló lépések:
  - Importálás táblázat kezelő szoftverbe (MS Excel, LO Calc)
  - Felesleges sorok és oszlopok eltávolítása
  - Cellaformátumok beállítása, tizedespont vs. tizedesvessző
  - Adatok egységesítése (pl. eltérő valuták egységessé alakítása)
  - Adatok aggregálása (pl. eltérő adatbázis-kezelők, OS-ek)
  - Táblázat exportálása célformátumba

Adatok tisztítása manuálisan

- A kiinduló adathalmazunk:

	A	B	C	D	E	F	G	H	I	J	K
1	TPC-C BENCHMARK RESULTS										
2	These results are valid as of date 6/12/2012 10:04:24 PM										
3											
4			TPC-C Results - Revision 5.X								
5											
6	Company	System	Spec. Revision	tpmC	Price/Perf	Total Sys. Cost	Currency	Database Software	Operating System	TP Monitor	Server CPU Type
7	Acer	Altos R710	5.5	66543	12.42	826507.55	AUD	Microsoft SQL Server	Microsoft Windows Serv	Microsoft CO	Intel Xeon - 3.6 GHz
8	Bull	Bull Escal	5.9	6085166	2.81	17127928	USD	IBM DB2 9.5	IBM AIX 5L V5.3	Microsoft CO	IBM POWER6 - 5.0
9	Bull	Bull Escal	5.9	629159	2.49	1566664	USD	IBM DB2 9.5 Enterpri	IBM AIX 5L V5.3	Microsoft CO	IBM POWER6 - 4.2
10	Bull	Bull Escal	5.8	1616162	3.54	5716286	USD	IBM DB2 9.1	IBM AIX 5L V5.3	Microsoft CO	IBM POWER6 - 4.7
11	Bull	Bull Escal	5.8	404462	3.51	1417121	USD	Oracle Database 10g	IBM AIX 5L V5.3	Microsoft CO	IBM POWER6 - 4.7

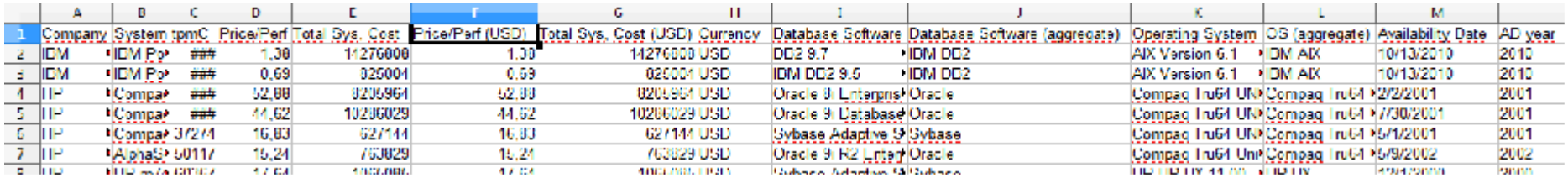
- Felesleges adatok:
  - Sorok (pl. a kezdő sorok, és az állomány végén lévő sorok, amelyek nem kapcsolódnak az eredményekhez)
  - Oszlopok (pl. Server CPU Type nekünk most nem kell)

Adatok tisztítása manuálisan

	A	B	C	D	E	F	G	H	I	J	K
1	TPC-C BENCHMARK RESULTS										
2	These results are valid as of date 6/12/2012 10:04:24 PM										
3											
4			TPC-C Results - Revision 5.X								
5											
6	Company	System	Spec. Revision	tpmC	Price/Perf	Total Sys. Cost	Currency	Database Software	Operating System	TP Monitor	Server CPU Type
7	Acer	Altos R710	5.5	66543	12.42	826507.55	AUD	Microsoft SQL Server	Microsoft Windows Serv	Microsoft CO	Intel Xeon - 3.6 GHz
8	Bull	Bull Escal	5.9	6085166	2.81	17127928	USD	IBM DB2 9.5	IBM AIX 5L V5.3	Microsoft CO	IBM POWER6 - 5.0
9	Bull	Bull Escal	5.9	629159	2.49	1566664	USD	IBM DB2 9.5 Enterpri	IBM AIX 5L V5.3	Microsoft CO	IBM POWER6 - 4.2
10	Bull	Bull Escal	5.8	1616162	3.54	5716286	USD	IBM DB2 9.1	IBM AIX 5L V5.3	Microsoft CO	IBM POWER6 - 4.7
11	Bull	Bull Escal	5.8	404462	3.51	1417121	USD	Oracle Database 10g	IBM AIX 5L V5.3	Microsoft CO	IBM POWER6 - 4.7

- További problémák:
  - Tizedesvessző vs. Pont
  - Eltérő valutákban megadott költségek
  - Ezeket cseréljük, konvertáljuk

Adatok tisztítása manuálisan



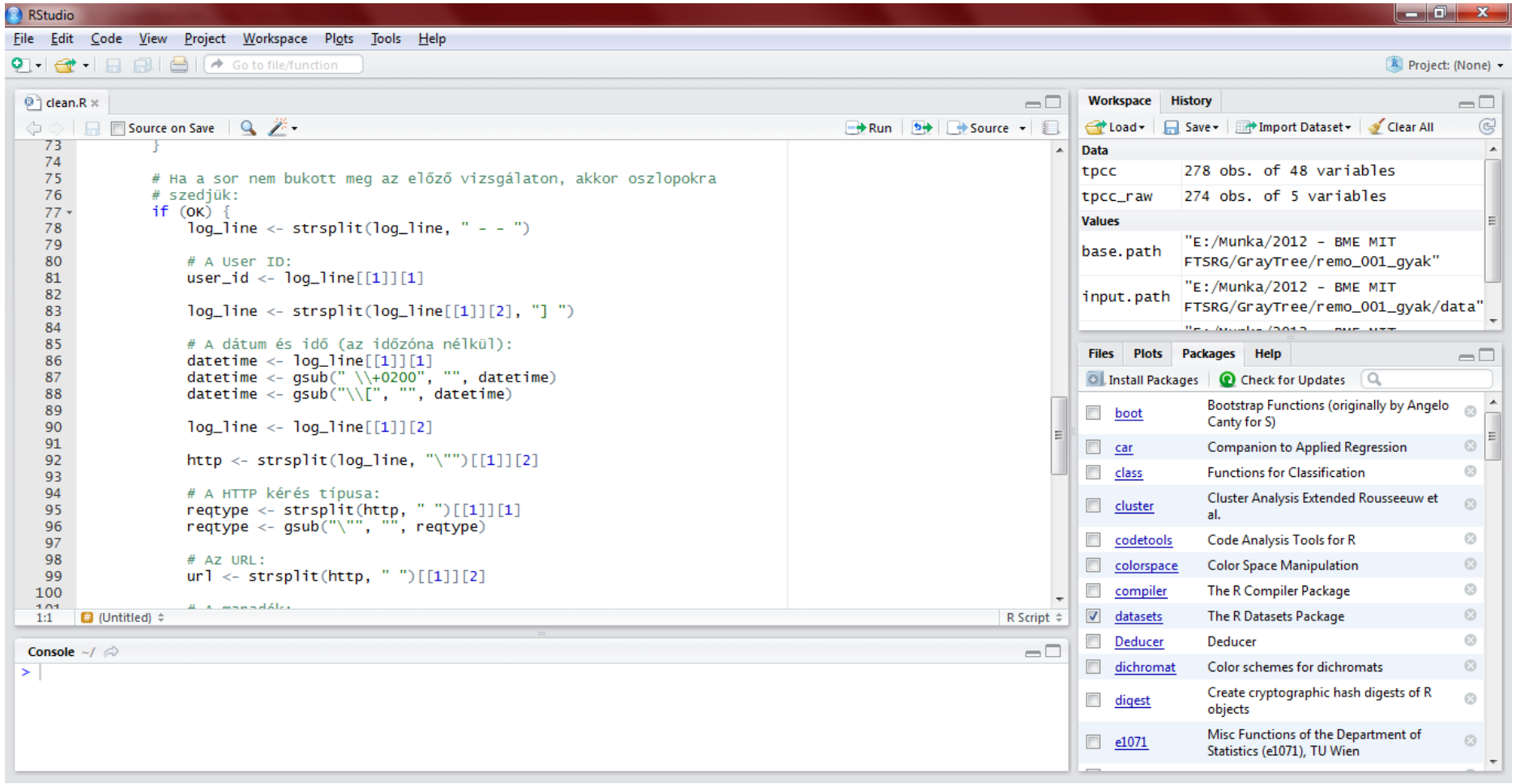
	A	B	C	D	E	F	G	H	I	J	K	L	M	
1	Company	System	tpmC	Price/Perf	Total Sys. Cost	Price/Perf (USD)	Total Sys. Cost (USD)	Currency	Database Software	Database Software (aggregate)	Operating System	OS (aggregate)	Availability Date	AD year
2	IBM	IBM Pe	###	1,36	11276000	1,36	11276000	USD	DB2 9.7	IBM DB2	AIX Version 6.1	IBM AIX	10/13/2010	2010
3	IBM	IBM Pe	###	0,69	925001	0,69	925001	USD	IBM DB2 9.6	IBM DB2	AIX Version 6.1	IBM AIX	10/13/2010	2010
4	HP	Compa	###	52,88	8200961	52,88	8200961	USD	Oracle 9i Enterprise	Oracle	Compaq Tru64 UNI	Compaq Tru64 UNI	2/2/2001	2001
5	HP	Compa	###	44,62	10200029	44,62	10200029	USD	Oracle 9i Database	Oracle	Compaq Tru64 UNI	Compaq Tru64 UNI	7/30/2001	2001
6	HP	Compa	37271	16,83	627144	16,83	627144	USD	Sybase Adaptive S	Sybase	Compaq Tru64 UNI	Compaq Tru64 UNI	6/1/2001	2001
7	HP	AlphaS	50117	15,24	763029	15,24	763029	USD	Oracle 9i R2 Enter	Oracle	Compaq Tru64 Unix	Compaq Tru64 UNI	6/9/2002	2002
8	HP	HP rx4	###	17,64	1053000	17,64	1053000	USD	Sybase Adaptive S	Sybase	HP UX 11.00	HP UX	12/31/2003	2003

- A tisztított adathalmazunkban:
  - Megfelelő adatformátumok
  - Aggregált értékek
    - Pl. nem foglakozunk az OS-ek és DBMS-ek verziójával (Windows Server 2003 és 2008 egységesen Windows-ként fog szerepelni)

Adatok tisztítása automatikusan

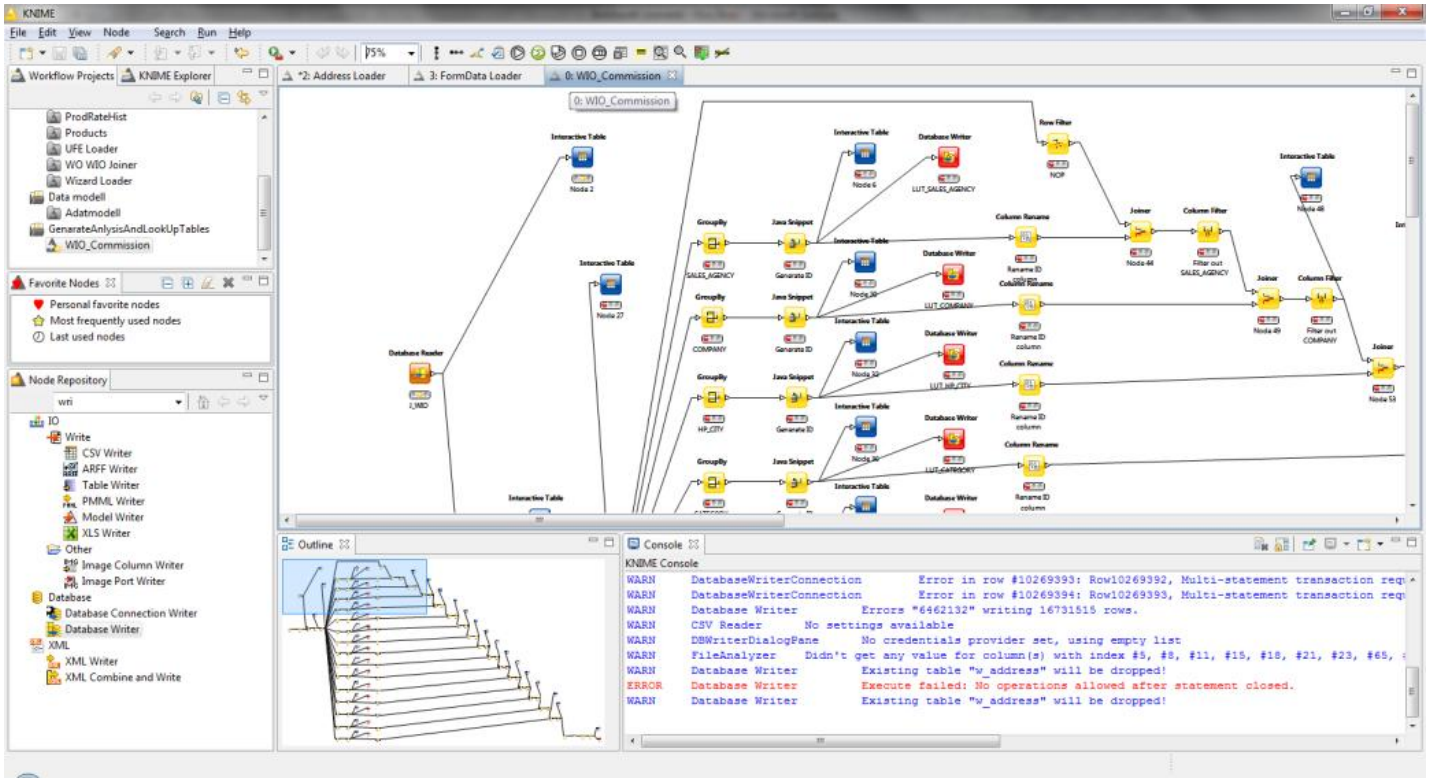
- Nagy méretű adathalmazok esetén nem alkalmazható a manuális adattisztítás
- Ilyenkor céleszközöket használunk:
  - Programozási vagy scriptnyelvek (pl. R, Perl, Python)
  - Grafikusan tervezett adatfeldolgozási folyamatok (pl. KNIME)
  - Későbbiekben látunk részletesebb példát (logelemzés)

Adatok tisztítása automatikusan - R



Adatok tisztítása automatikusan - KNIME

- Adattisztítás + riportgenerálás + adatelemzési eszközök elérése



VIZUÁLIS ELEMZÉS

Cél: tisztított adat vizuális elemzése

Elemzési kérdések

- Mely években megjelent konfigurációkat tartalmazza a benchmark? Mennyire használható/releváns napjainkban?
- Az egyes beszállítók mely években voltak aktívak? Mely beszállító a nagy játékos?
- Ha cégünk igényeit egy alacsonyabb teljesítményű konfigurációval is ki tudjuk elégíteni, akkor mely beszállítók közül válasszunk?
- Mit lehet megállapítani a teljesítmény változásáról?
- Hogyan alakulnak a tranzakciós és összköltségek?
- Milyen adatbázis-kezelő szoftvert válasszunk, ha még mindig az olcsóbb megoldást szeretnénk használni?
- Milyen operációs rendszert válasszunk, ha a teljes költséget minimalizálni akarjuk?
- Melyik beszállító melyik adatbázis-kezelőben és operációs rendszerben ,,utazik''?
- Melyik beszállítót, adatbázis-kezelőt és operációs rendszert tartalmazza a leggyakrabban az adathalmaz?
- Próbáljunk összefüggést találni a teljesítmények, költségek, operációs rendszerek és adatbázis-kezelők között!

GYAKORLAT (MONDRIAN)