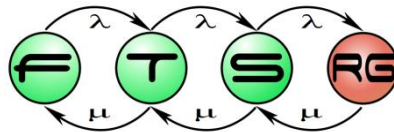


Vizuális adatelemzés - Gyakorlat



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

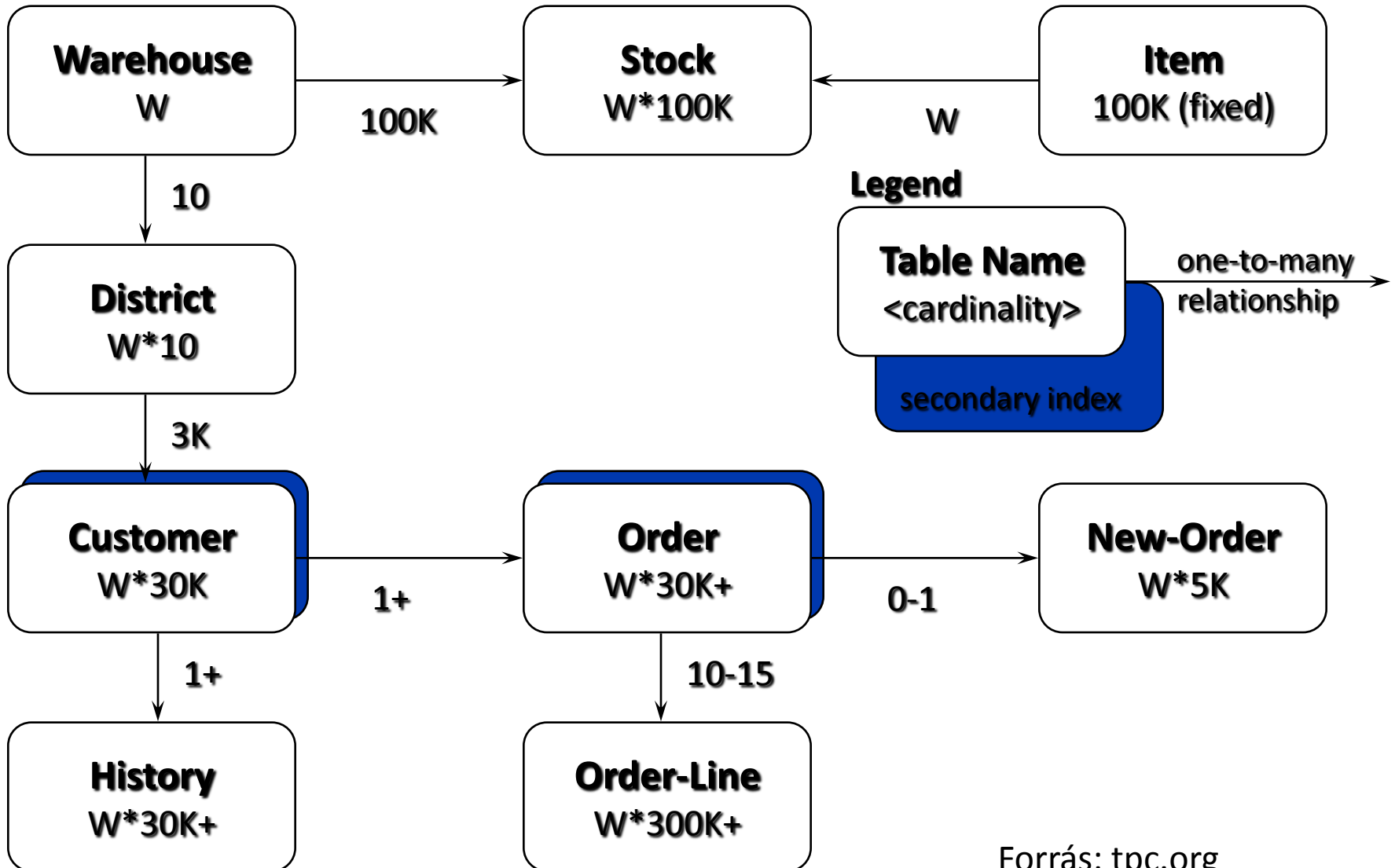
BEMUTATÓ (MONDRIAN)

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	<u>Company</u>	<u>System</u>	<u>Spec. Revision</u>	<u>tpmC</u>	<u>Price/Perf</u>	<u>Total Sys. Cost</u>	<u>Currency</u>	<u>Database Software</u>	<u>Operating System</u>	<u>TP Monitor</u>	<u>Server CPU Type</u>
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 Enterprise	▶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	<u>Company</u>	<u>System</u>	<u>Spec. Revision</u>	<u>tpmC</u>	<u>Price/Perf</u>	<u>Total Sys. Cost</u>	<u>Currency</u>	<u>Database Software</u>	<u>Operating System</u>	<u>TP Monitor</u>	<u>Server CPU Type</u>
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 10q	▶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

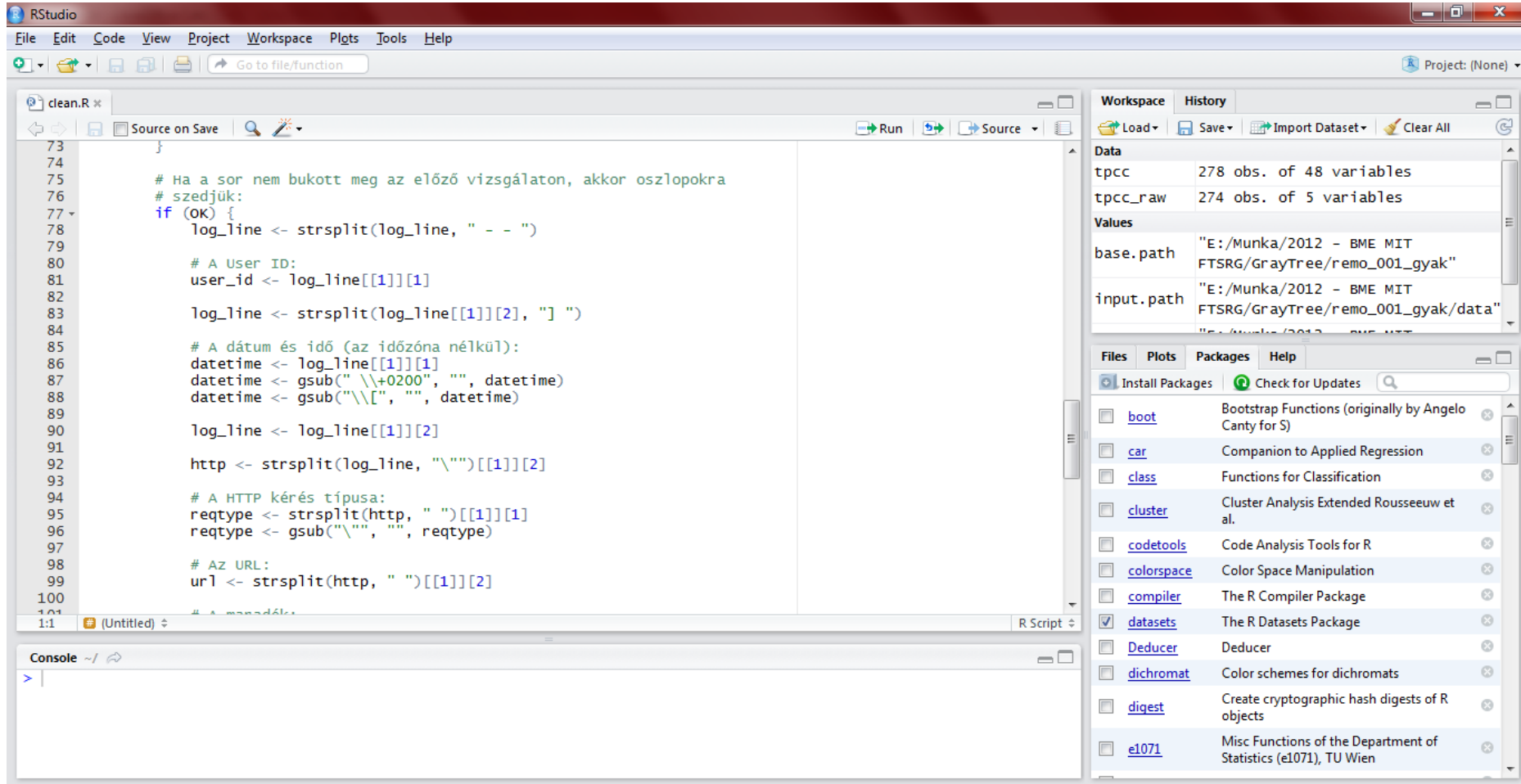
	A	B	C	D	E	F	G	H	I	J	K	L	M	
1	Company	System	pmC	Price/Perf	Total Sys. Cost	Price/Perf (USD)	Total Sys. Cost (USD)	Currency	Database Software	Database Software (aggregate)	Operating System	OS (aggregate)	Availability Date	AD year
2	IDM	IDM Pp	###	1.38	14276000	1.38	14276000	USD	DC2 9.7	IDM DC2	AIX Version 6.1	DM AIX	10/13/2010	2010
3	IDM	IDM Pp	###	0.69	820004	0.69	820004	USD	IDM DC2 9.5	IDM DC2	AIX Version 6.1	DM AIX	10/13/2010	2010
4	HP	Compa	###	52.88	8206964	52.88	8206964	USD	Oracle 9i Enterprise	Oracle	Compaq Inru64 UN	Compaq Inru64	2/2/2001	2001
5	HP	Compa	###	44.62	10206029	44.62	10206029	USD	Oracle 9i Database	Oracle	Compaq Inru64 UN	Compaq Inru64	7/30/2001	2001
6	HP	Compa	3/2/1	16.83	627144	16.83	627144	USD	Sybase Adaptive S	Sybase	Compaq Inru64 UN	Compaq Inru64	5/1/2001	2001
7	HP	Alpha	50117	15.24	763829	15.24	763829	USD	Oracle 9i R2 Enterprise	Oracle	Compaq Inru64 Uni	Compaq Inru64	5/9/2002	2002
8	HP	HP	HP	47.84	4064881	47.84	4064881	USD	Subsee Adaptive S	Subsee	HP	HP	4/14/2001	2001

- A tisztított adathalmazunkban:
 - Megfelelő adatformátumok
 - Aggregált értékek
 - Pl. nem foglalkozunk 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



The screenshot displays the RStudio interface. The main editor window shows an R script named 'clean.R' with the following code:

```
73 }  
74  
75 # Ha a sor nem bukott meg az előző vizsgálaton, akkor oszlopokra  
76 # szedjük:  
77 if (OK) {  
78   log_line <- strsplit(log_line, " - - ")  
79  
80   # A User ID:  
81   user_id <- log_line[[1]][1]  
82  
83   log_line <- strsplit(log_line[[1]][2], " ")  
84  
85   # A dátum és idő (az időzóna nélküli):  
86   datetime <- log_line[[1]][1]  
87   datetime <- gsub("\\+0200", "", datetime)  
88   datetime <- gsub("\\[", "", datetime)  
89  
90   log_line <- log_line[[1]][2]  
91  
92   http <- strsplit(log_line, "\\\"")[[1]][2]  
93  
94   # A HTTP kérés típusa:  
95   reqtype <- strsplit(http, " ")[1][1]  
96   reqtype <- gsub("\\\"", "", reqtype)  
97  
98   # Az URL:  
99   url <- strsplit(http, " ")[1][2]  
100  
101 # A maradék
```

The right-hand side of the interface shows the 'Workspace' and 'History' panels. The 'Data' section lists:

- tpcc: 278 obs. of 48 variables
- tpcc_raw: 274 obs. of 5 variables

The 'Values' section shows:

- base.path: "E:/Munka/2012 - BME MIT FTSRG/GrayTree/remo_001_gyak"
- input.path: "E:/Munka/2012 - BME MIT FTSRG/GrayTree/remo_001_gyak/data"

The 'Files' panel shows a search bar and a list of installed packages:

- boot: Bootstrap Functions (originally by Angelo Canty for S)
- car: Companion to Applied Regression
- class: Functions for Classification
- cluster: Cluster Analysis Extended Rousseeuw et al.
- codetools: Code Analysis Tools for R
- colorspace: Color Space Manipulation
- compiler: The R Compiler Package
- datasets: The R Datasets Package
- Deducer: Deducer
- dichromat: Color schemes for dichromats
- digest: Create cryptographic hash digests of R objects
- e1071: Misc Functions of the Department of Statistics (e1071), TU Wien

The console at the bottom shows the prompt '> |'.

Adatok tisztítása automatikusan - KNIME

- Adattisztítás + riportgenerálás + adatelemzési eszközök elérése

The screenshot displays the KNIME software interface with a workflow titled "WIO_Commission". The workflow consists of several interconnected nodes: Database Reader, Interactive Table, GroupBy, Java Snippet, Interactive Table, Database Writer, Column Rename, Row Filter, Joiner, Column Filter, and another Interactive Table. The nodes are arranged in a hierarchical structure, with data flowing from left to right. The console at the bottom shows the following error messages:

```
KNIME Console
WARN DatabaseWriterConnection Error in row #10269393: Row10269392, Multi-statement transaction req
WARN DatabaseWriterConnection Error in row #10269394: Row10269393, Multi-statement transaction req
WARN Database Writer Errors "6462132" writing 16791515 rows.
WARN CSV Reader No settings available
WARN DBWriterDialogPane No credentials provider set, using empty list
WARN FileAnalyzer Didn't get any value for column(s) with index #5, #8, #11, #15, #18, #21, #23, #45,
WARN Database Writer Existing table "w_address" will be dropped!
ERROR Database Writer Execute failed: No operations allowed after statement closed.
WARN Database Writer Existing table "w_address" will be dropped!
```

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)