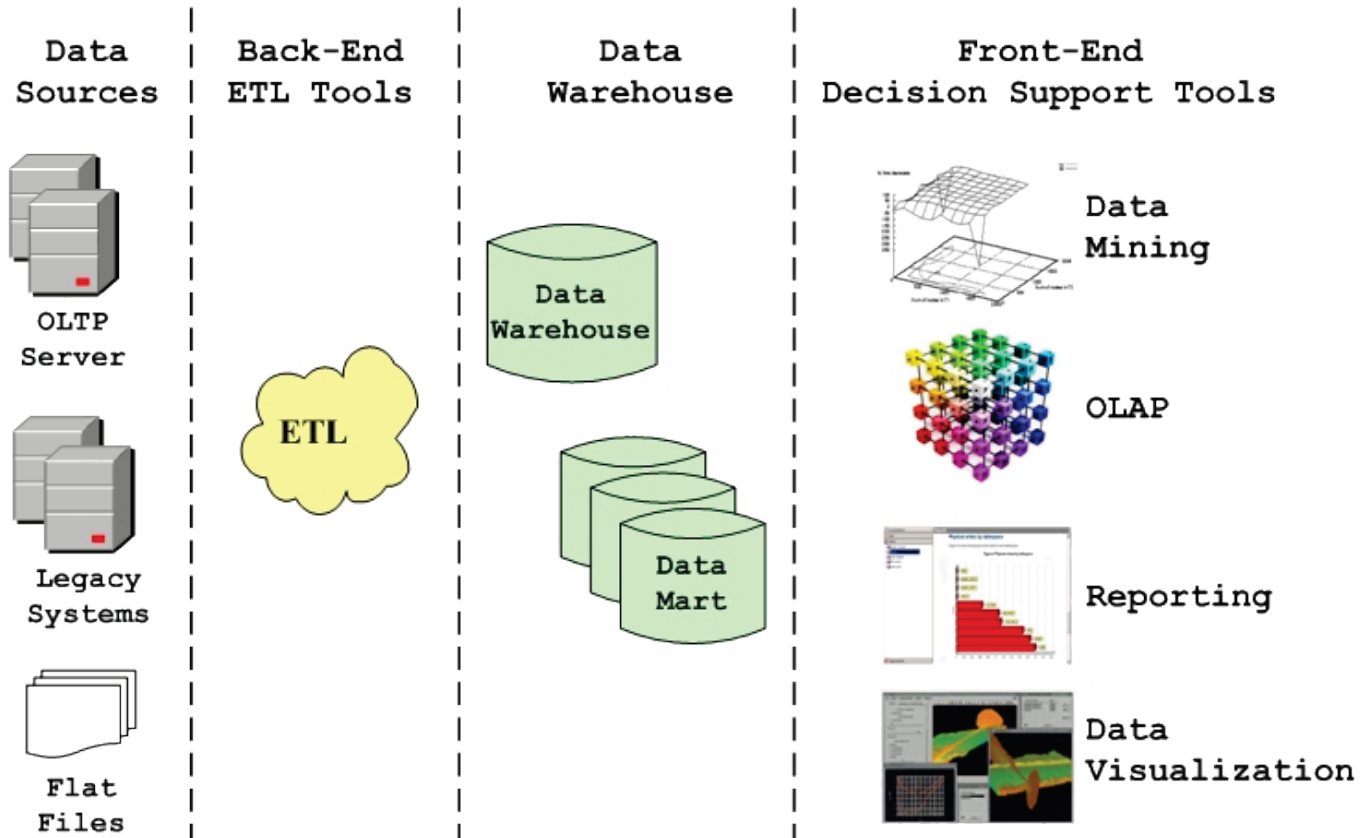


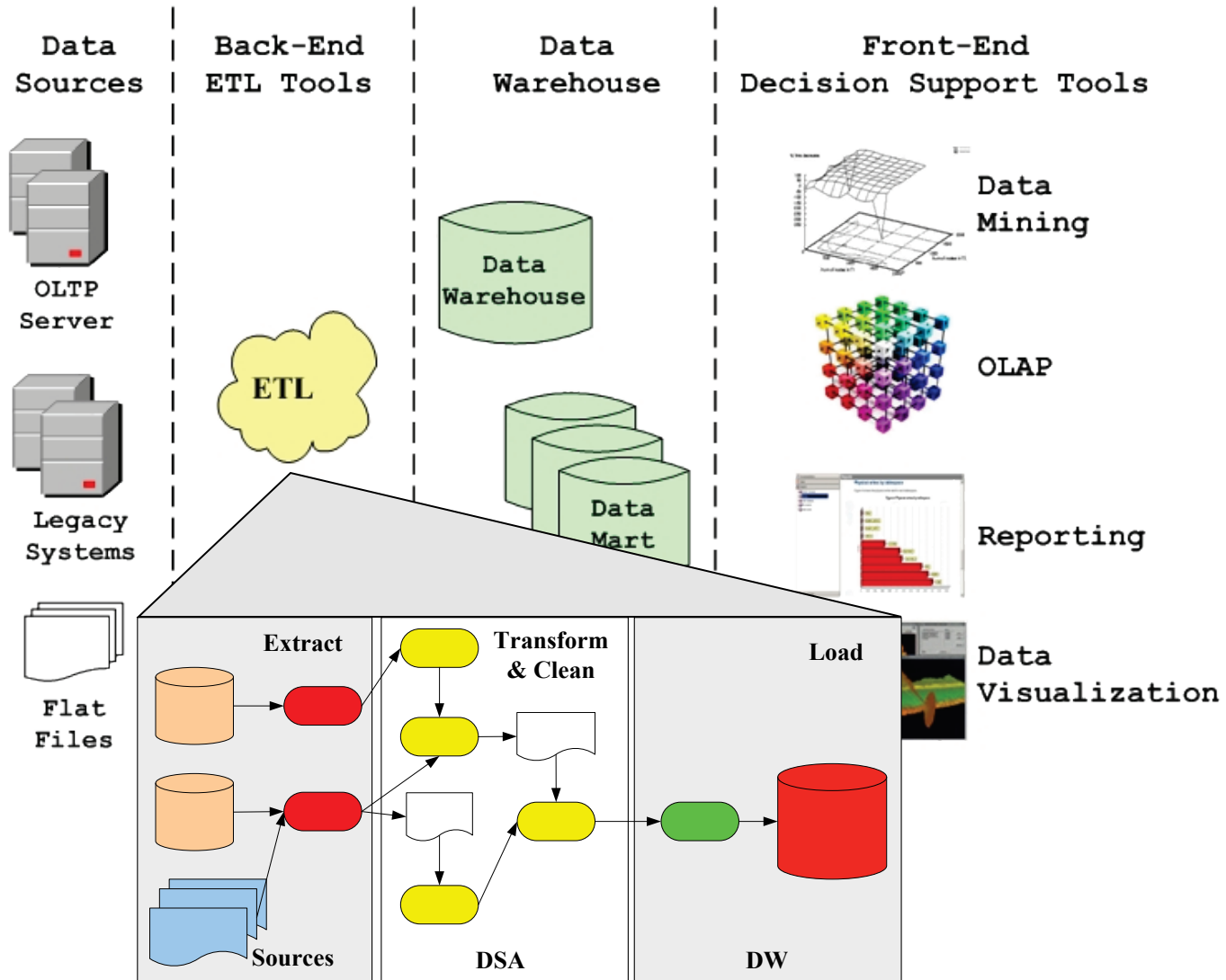
Data Warehouse Refreshment via ETL tools

Panos Vassiliadis

Data Warehouse Environment



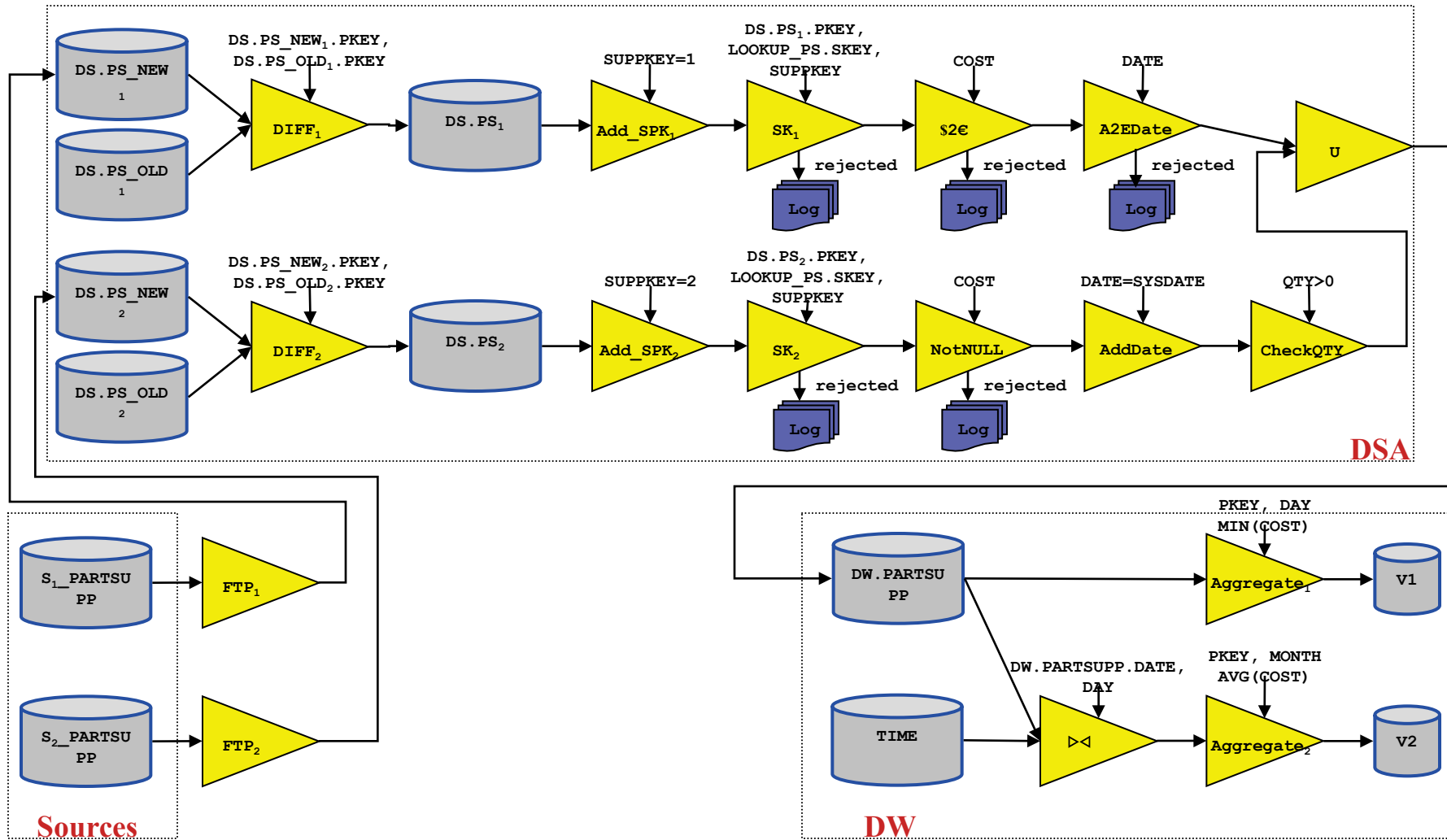
Extract-Transform-Load (ETL)



Importance

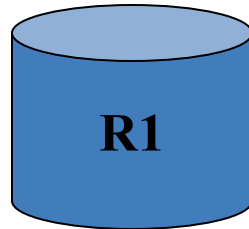
- ETL market has a steady increase rate of approximately 20.1% each year, while it becomes a \$667 million market in 2001 (*Giga'02*)
- ETL and Data Cleaning tools **cost**
 - **30%** of effort and expenses in the budget of the DW (*Enterprise Information Portals*)
 - **55%** of the total costs of DW runtime (*Inmon*)
 - **80%** of the development time in a DW project (*Demarest*)
- ETL tools will not be replaced by EAI (Enterprise Application Integration) tools in near future (*Giga'02*)
- ETL tools will be used in other areas beyond DWs (*Gartner'04*)

ETL workflows

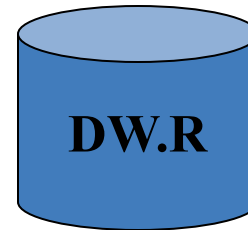
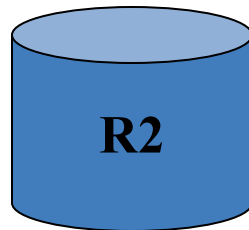


Value Incompatibility (example of surrogate keys)

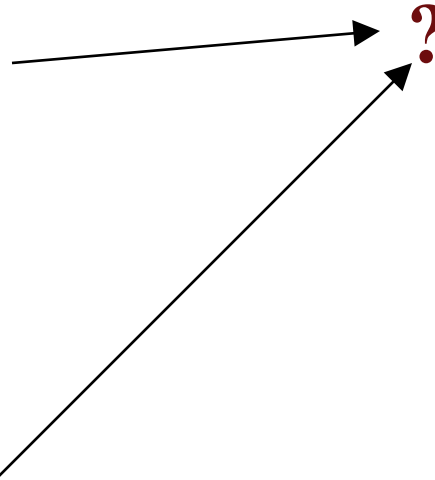
<u>ID</u>	Descr
10	Coke
20	Pepsi



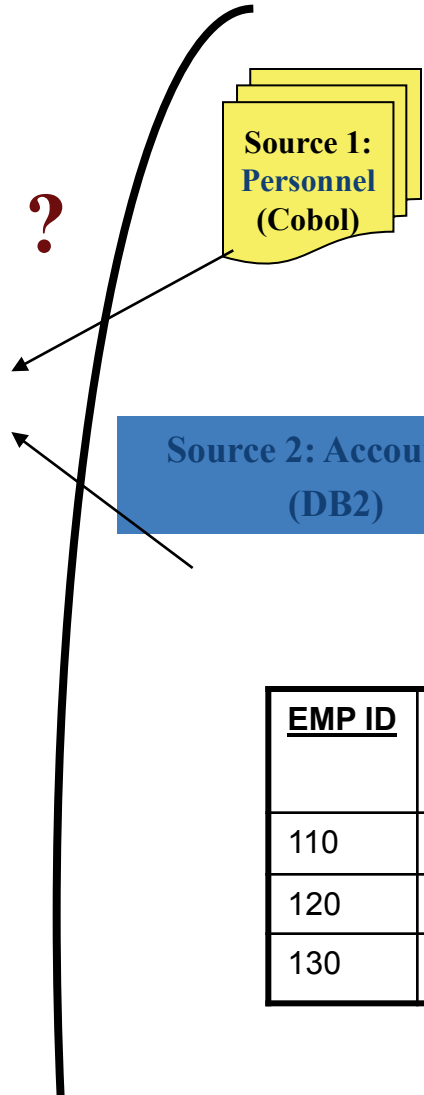
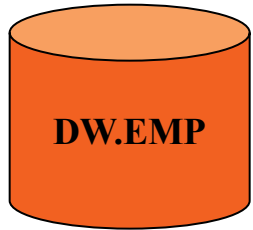
<u>ID</u>	Descr
10	Pepsi
20	Fanta



<u>ID</u>	Descr
??	??
??	??



Data mappings

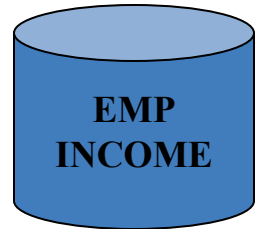


Source 1:
Personnel
(Cobol)

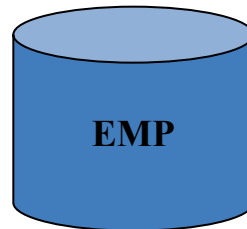
<u>EMP ID</u>	Name	DoB	Salary	Total Income	DeptID
110	Kostas	1/1/72	1500	1200	132
...

Source 2: Accounting
(DB2)

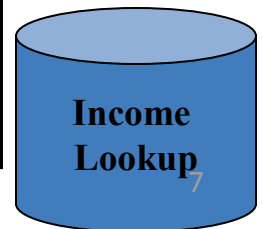
<u>EMP ID</u>	<u>IL ID</u>	Amount
110	10	1500
110	30	300



<u>EMP ID</u>	Name	Age
110	Kostas	30
120	Mitsos	48
130	Roula	29



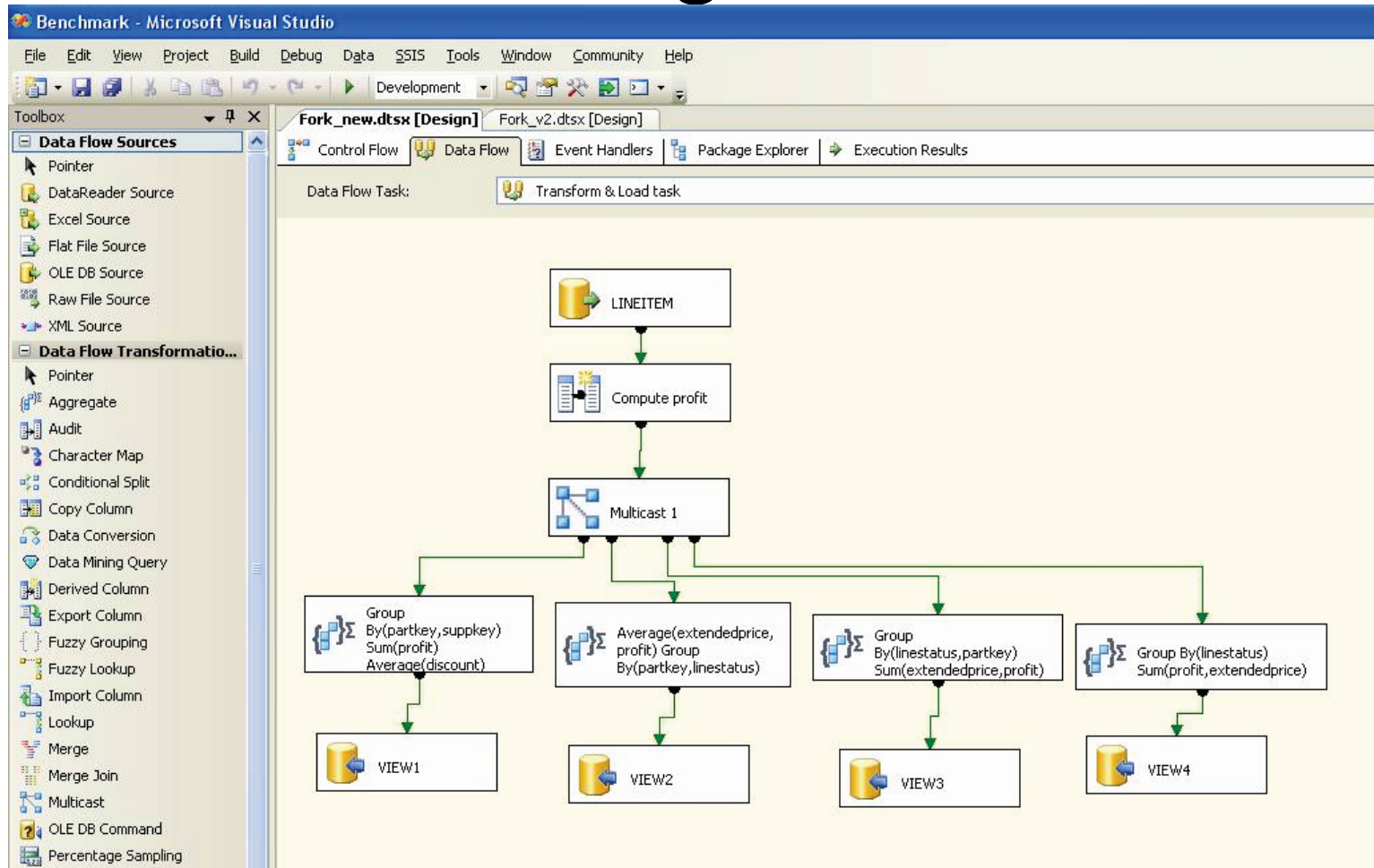
<u>IL ID</u>	Descr
10	Salary
20	Bonus 1
30	Tax
...	...



DW

MS SSIS

SQL Server Integration Services



Talend Open Studio for Data Integration

www.talend.com/download_form.php?cont=gen&src=HomePage

The screenshot displays the Talend Open Studio interface. The main workspace shows a job design with the following components and connections:

- US_States** (Source) connects to **row2 (lookup)**.
- Customers** (Source) connects to **row1 (Main)**.
- row1 (Main)** and **row2 (lookup)** both connect to **tMap_1**.
- tMap_1** has three output paths:
 - New_Customers (Main order:1)** connects to **Target_Customers_Table**.
 - Virginia_customers (Main order:2)** connects to **Target_Virginia_Customers_Table**.
 - Rejected_Customers (Main order:3)** connects to **tFileOutputExcel_1**.

The **Target_Customers_Table (MysqlOutput_1)** configuration panel is visible, showing the following settings:

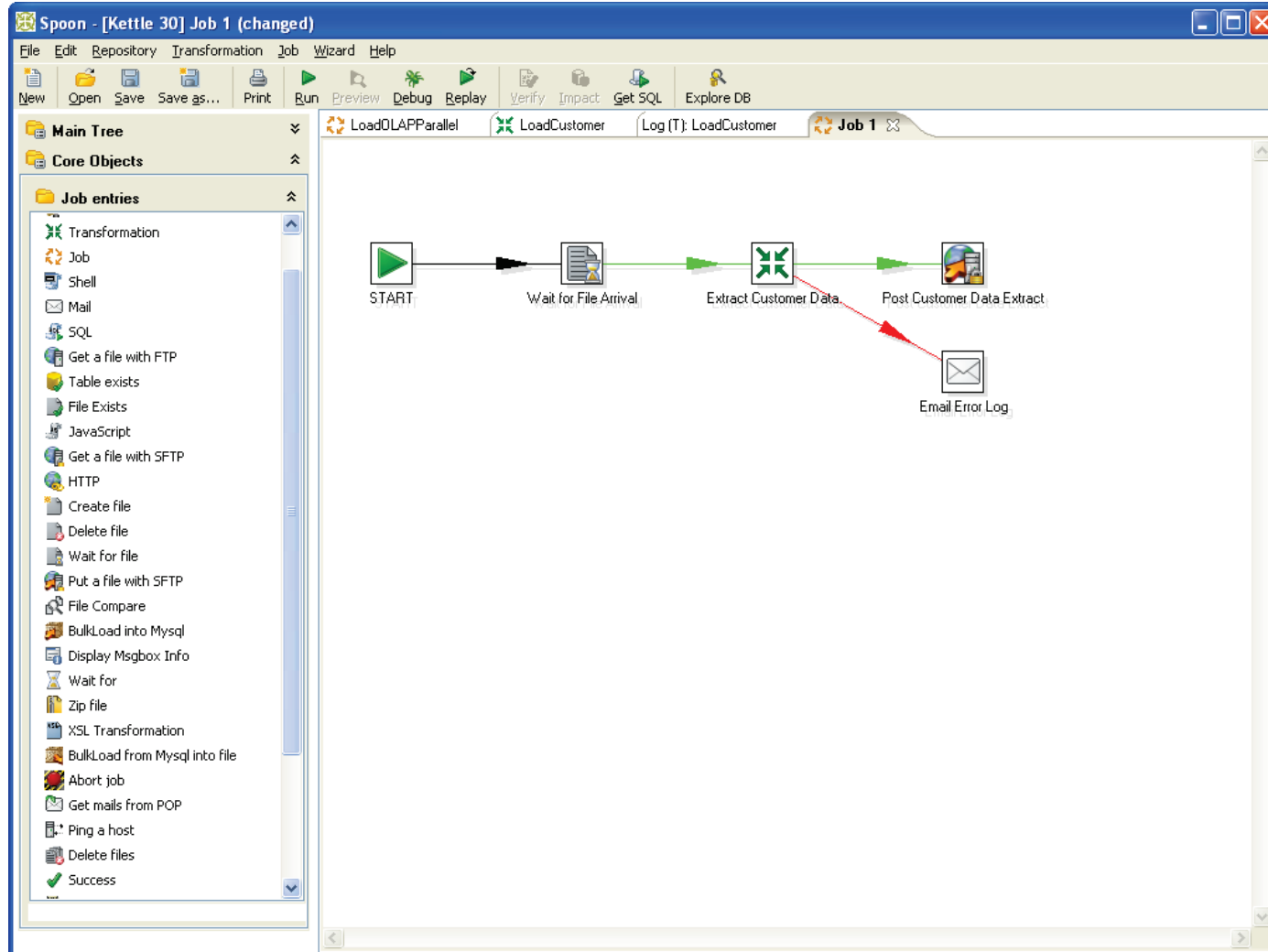
Property	Type	Value
Repository	Repository	DB (MYSQL):demoMysql
DB Version	DB Version	Mysql 5
Use an existing connection	Boolean	<input type="checkbox"/>
Host	Host	localhost
Port	Port	3306
Database	Database	demoproject
Username	Username	root
Password	Password	*****
Table	Table	customer
Action on table	Action on table	Default
Action on data	Action on data	Insert
Schema	Schema	Repository DB (MYSQL):demoMysql - customer
Die on error	Boolean	<input type="checkbox"/>

The **Code Viewer** shows the following code snippet:

```
/**  
 * [tMysqlOutput_1 main ] :  
 */  
  
currentComponent="tMysqlOut;  
  
whetherReject_tMysqlOutput_  
false;  
  
if (New_Customers.idcustomer  
null) {
```

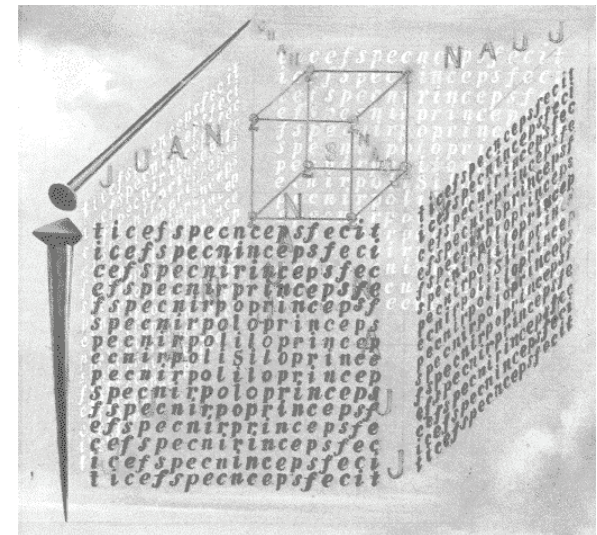
Pentaho's Kettle

<http://kettle.pentaho.com/>



OLAP & data cubes

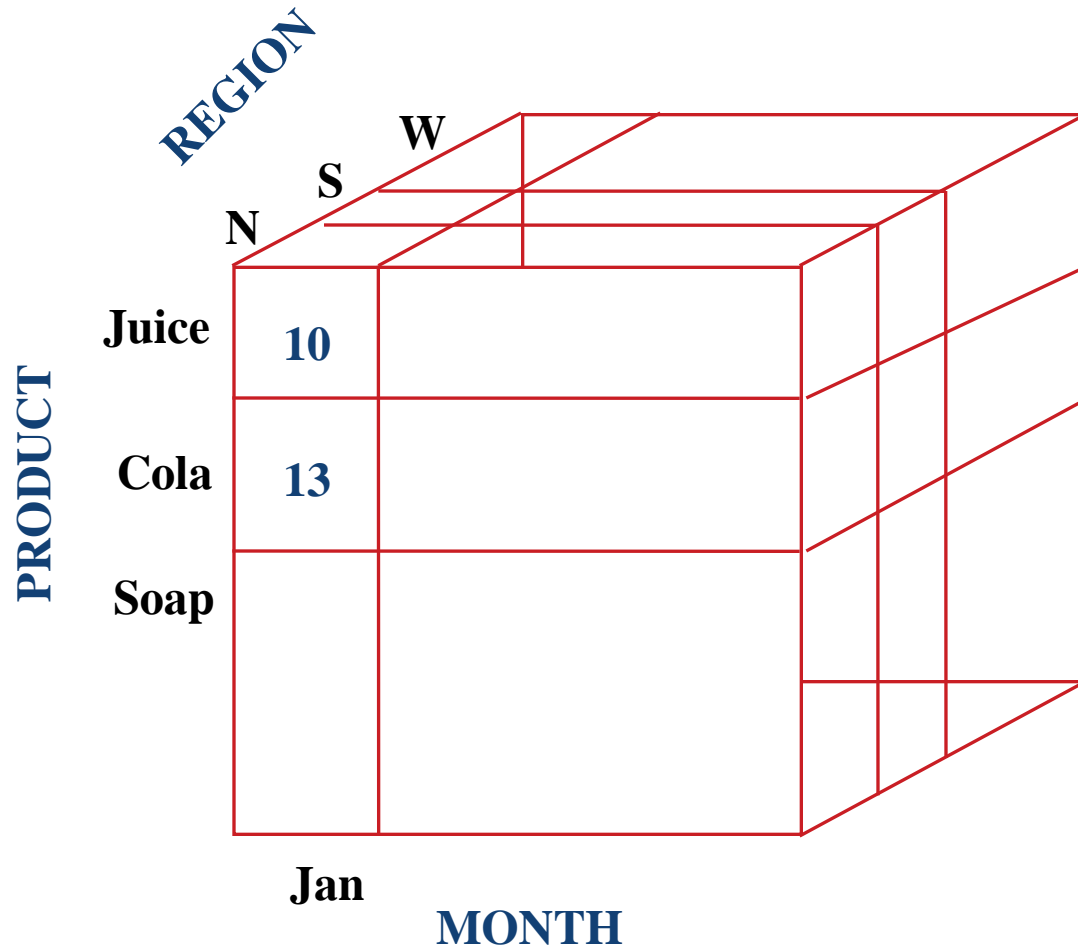
Panos Vassiliadis



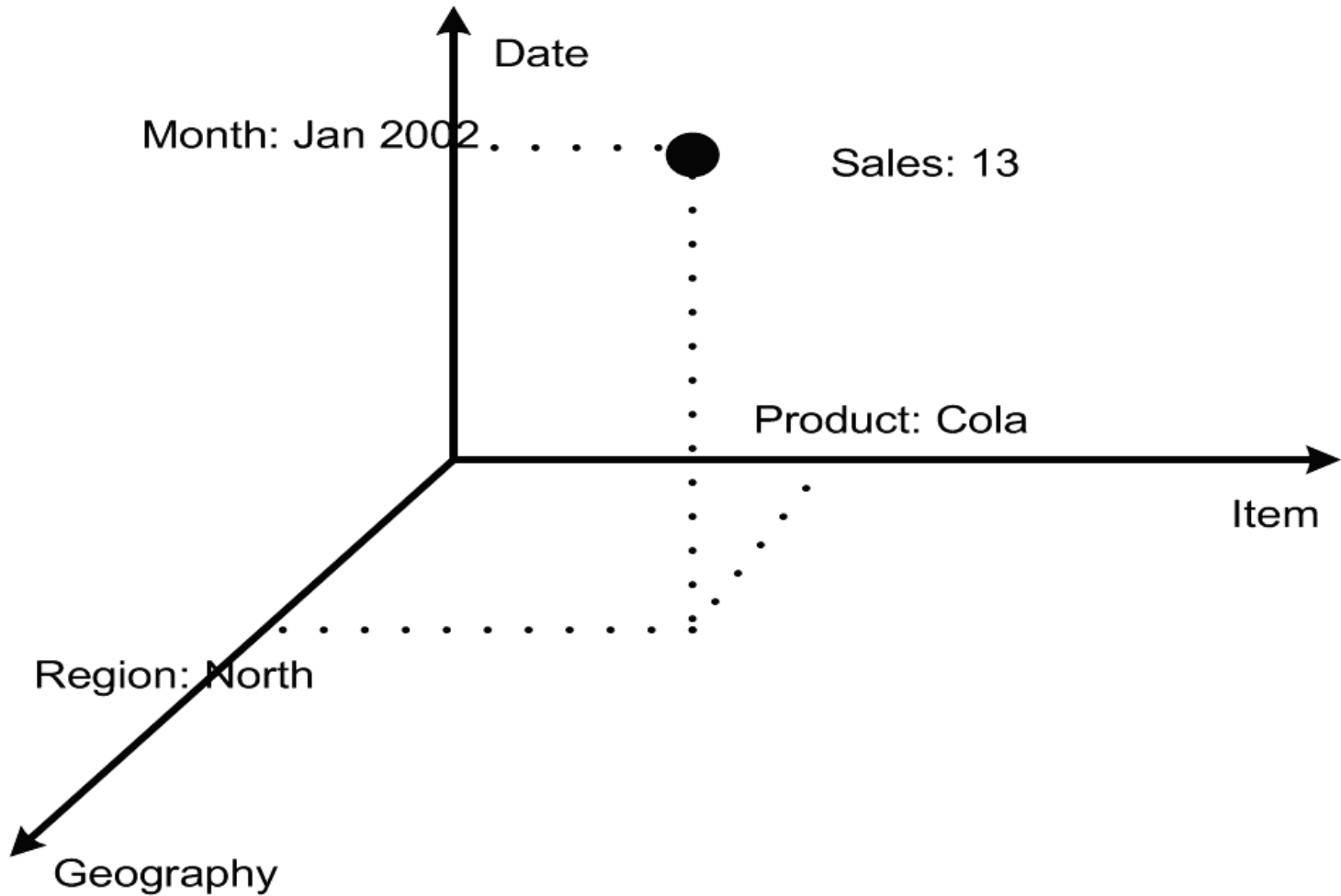
OLAP

- Αφορά την ανάλυση κάποιων μετρήσιμων μεγεθών (μέτρων)
 - πωλήσεις, απόθεμα, κέρδος,...
- Διαστάσεις: παράμετροι που καθορίζουν το περιβάλλον (context) των μέτρων
 - ημερομηνία, προϊόν, τοποθεσία, πωλητής, ...
- Κύβοι: συνδυασμοί διαστάσεων που καθορίζουν κάποια μέτρα
 - Ο κύβος καθορίζει ένα πολυδιάστατο χώρο διαστάσεων, με τα μέτρα να είναι σημεία του χώρου αυτού

Κύβοι για OLAP



Κύβοι για OLAP



Βασικές Έννοιες OLAP

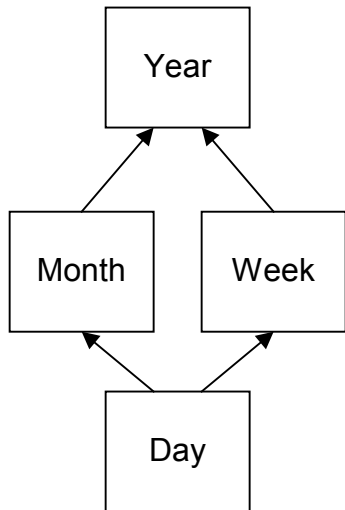
- Τα δεδομένα θεωρούνται αποθηκευμένα σε ένα **πολυδιάστατο πίνακα** (multi-dimensional array), ο οποίος αποκαλείται και **κύβος** ή **υπερκύβος** (**Cube** και **HyperCube** αντίστοιχα).
- Ο κύβος είναι μια ομάδα από **κελιά** δεδομένων (data cells). Κάθε κελί χαρακτηρίζεται μονοσήμαντα από τις αντίστοιχες τιμές των **διαστάσεων** (dimensions) του κύβου.
- Τα περιεχόμενα του κελιού ονομάζονται **μέτρα** (measures) και αναπαριστούν τις αποτιμώμενες αξίες του πραγματικού κόσμου.

Ιεραρχίες επιπέδων για OLAP

- Μια **διάσταση** μοντελοποιεί όλους τους τρόπους με τους οποίους τα δεδομένα μπορούν να συναθροιστούν σε σχέση με μια συγκεκριμένη παράμετρο του περιεχομένου τους.
 - Ημερομηνία, Προϊόν, Τοποθεσία, Πωλητής, ...
- Κάθε διάσταση έχει μια σχετική **ιεραρχία επιπέδων** συνάθροισης των δεδομένων (hierarchy of levels). Αυτό σημαίνει, ότι η διάσταση μπορεί να θεωρηθεί από πολλά επίπεδα αδρομέρειας.
 - Ημερομηνία: **μέρα, εβδομάδα, μήνας, χρόνος, ...**

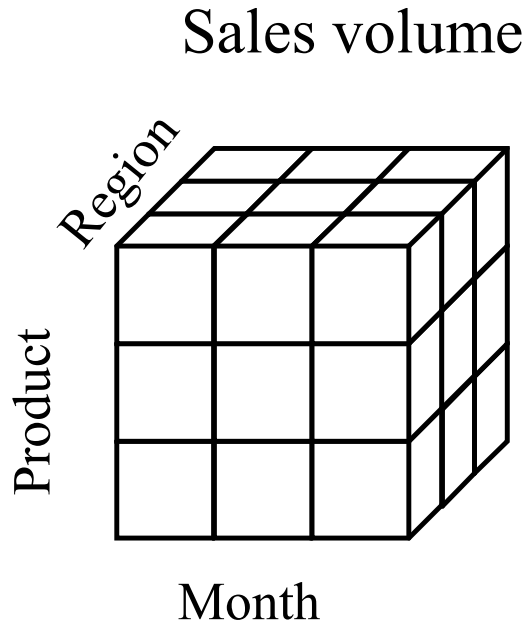
Ιεραρχίες Επιπέδων

- **Ιεραρχίες Επιπέδων:** κάθε διάσταση οργανώνεται σε διαφορετικά επίπεδα αδρομέρειας
- Ο χρήστης μπορεί να πλοηγηθεί από το ένα επίπεδο στο άλλο, δημιουργώντας νέους κύβους κάθε φορά



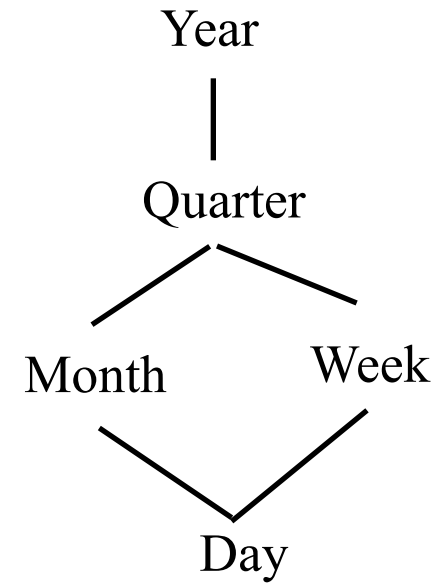
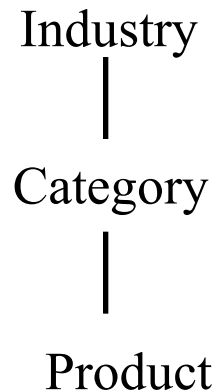
Αδρομέρεια: το αντίθετο της λεπτομέρειας
-- ο σωστός όρος είναι αδρομέρεια...

Κύβοι & ιεραρχίες διαστάσεων για OLAP



Διαστάσεις: Product, Region, Date

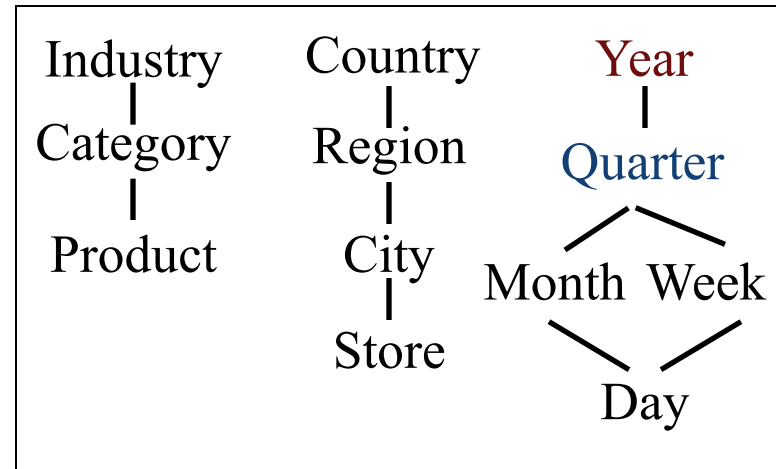
Ιεραρχίες διαστάσεων:



Εργασίες που κάνει ο χρήστης

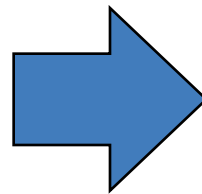
- Συνήθεις πράξεις που κάνουμε σε κύβους
 - Συναθροίσεις (total sales, percent-to-total)
 - Συγκρίσεις (budget vs. expense)
 - Ταξινόμηση - κατάταξη (top 10)
 - Πρόσβαση σε πιο αναλυτική πληροφορία
 - Οπτικοποίηση με διαφορετικούς τρόπους

Roll up



Sales volume			
	Products	Store1	Store2
Q1	Electronics	\$5,2	\$5,6
	Toys	\$1,9	\$1,4
	Clothing	\$2,3	\$2,6
	Cosmetics	\$1,1	\$1,1
Q2	Electronics	\$8,9	\$7,2
	Toys	\$0,75	\$0,4
	Clothing	\$4,6	\$4,6
	Cosmetics	\$1,5	\$0,5

Χρόνος: Επίπεδο **Quarter**



Sales volume			
	Products	Store1	Store2
Year 1996	Electronics	\$14,1	\$12,8
	Toys	\$2,65	\$1,8
	Clothing	\$6,9	\$7,2
	Cosmetics	\$2,6	\$1,6

Χρόνος: Επίπεδο **Year**

SUM(Sales volumes)