# Saiku – taking OLAP databases into 21st century

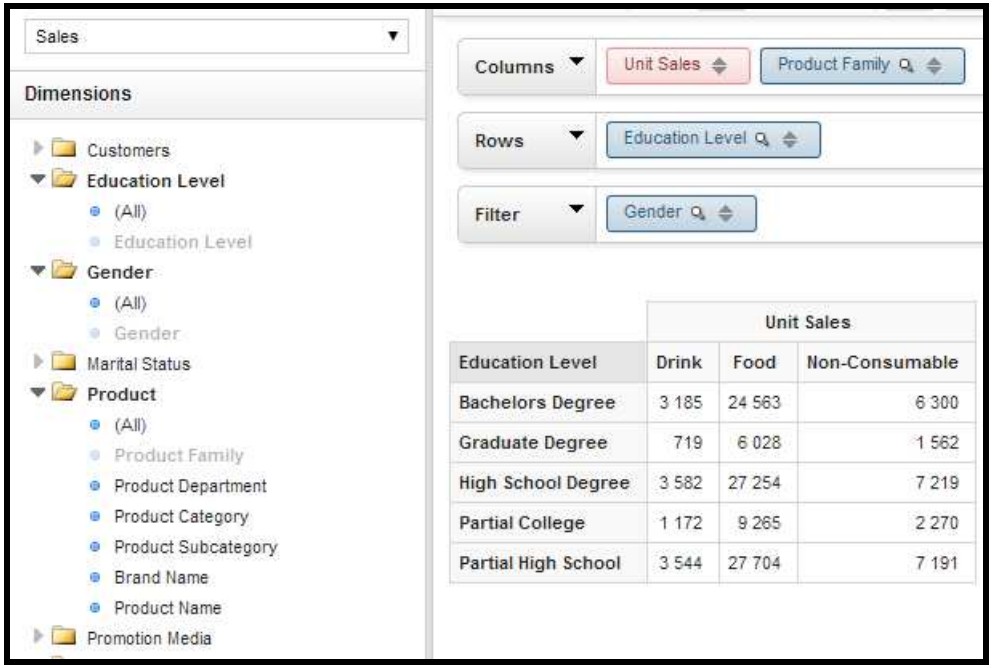

## Tomasz Nurkiewicz

nurkiewicz.com | @tnurkiewicz


Slides: bit.ly/33degree

# What is Saiku?

## DEMO



# Core concepts

- OLAP
- Fact
- Dimension
- Hierarchy

# Example facts

- Sold product
- Tweet/forum post/shared photo
- Website hit
- Incoming text message
- ...you name it

# Dimension

"Properties of facts"

- When?
- What?
- Where?
- Who?
- How?

# Example dimensions

## Access log

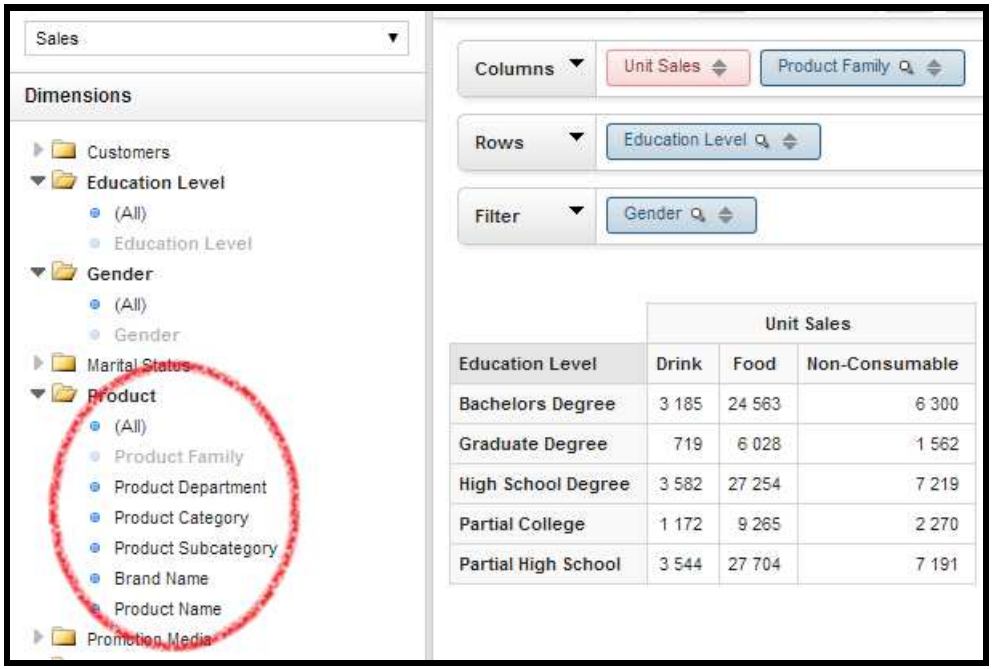
- Timestamp
- IP
- URL resource
- HTTP response code

# Hierarchy

## Multi-level aggregation

### Example: *location* hierarchy

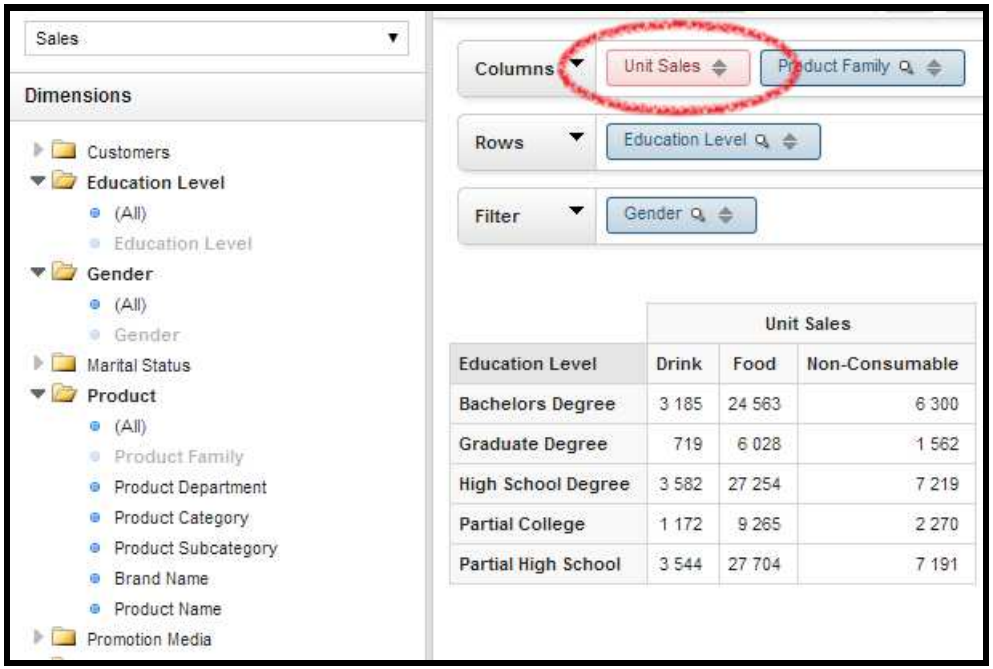
- (All)
- Continent
- Country
- State
- City




| | Unit Sales | | |
|---|---|---|---|
| Education Level | Drink | Food | Non-Consumable |
| Bachelors Degree | 3 185 | 24 563 | 6 300 |
| Graduate Degree | 719 | 6 028 | 1 562 |
| High School Degree | 3 582 | 27 254 | 7 219 |
| Partial College | 1 172 | 9 265 | 2 270 |
| Partial High School | 3 544 | 27 704 | 7 191 |

# Measures

- Quantitative properties
- Aggregate matching facts over them
- Count/Sum/Average/Min/Max

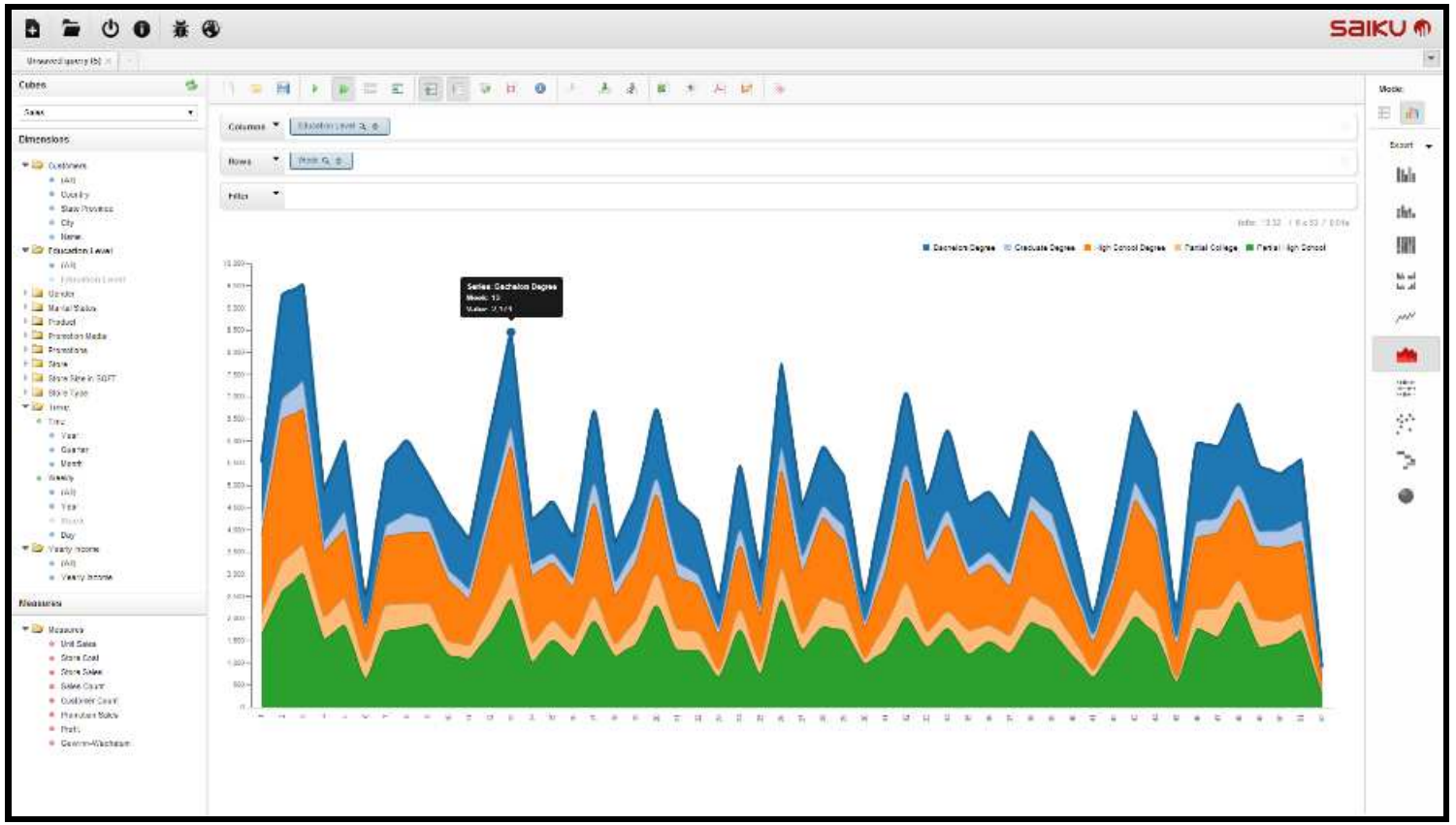



| | Unit Sales | | |
|---|---|---|---|
| Education Level | Drink | Food | Non-Consumable |
| Bachelors Degree | 3 185 | 24 563 | 6 300 |
| Graduate Degree | 719 | 6 028 | 1 562 |
| High School Degree | 3 582 | 27 254 | 7 219 |
| Partial College | 1 172 | 9 265 | 2 270 |
| Partial High School | 3 544 | 27 704 | 7 191 |

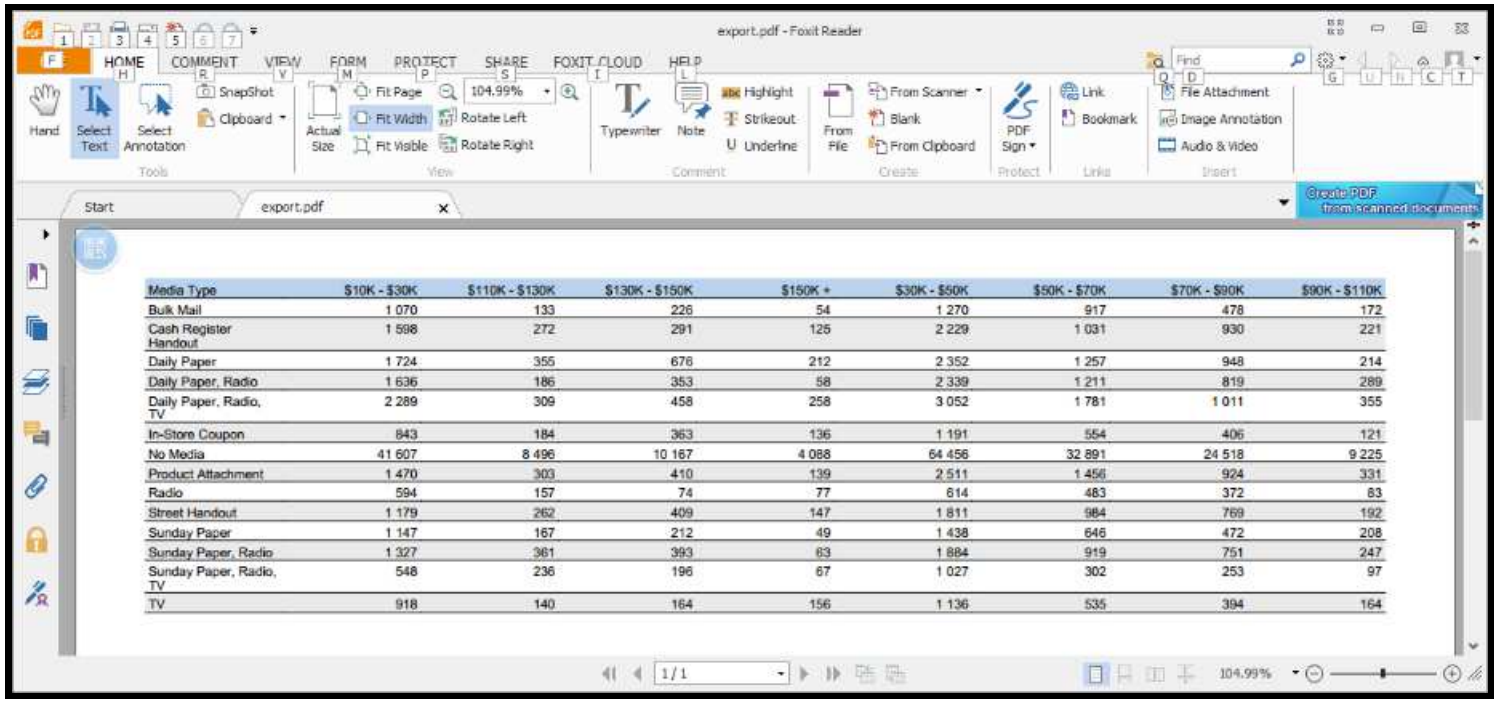
# Example measures

- Load time (*page hit fact*)
- Total price (*sale fact*)
- Age of customer

# Charting - DEMO

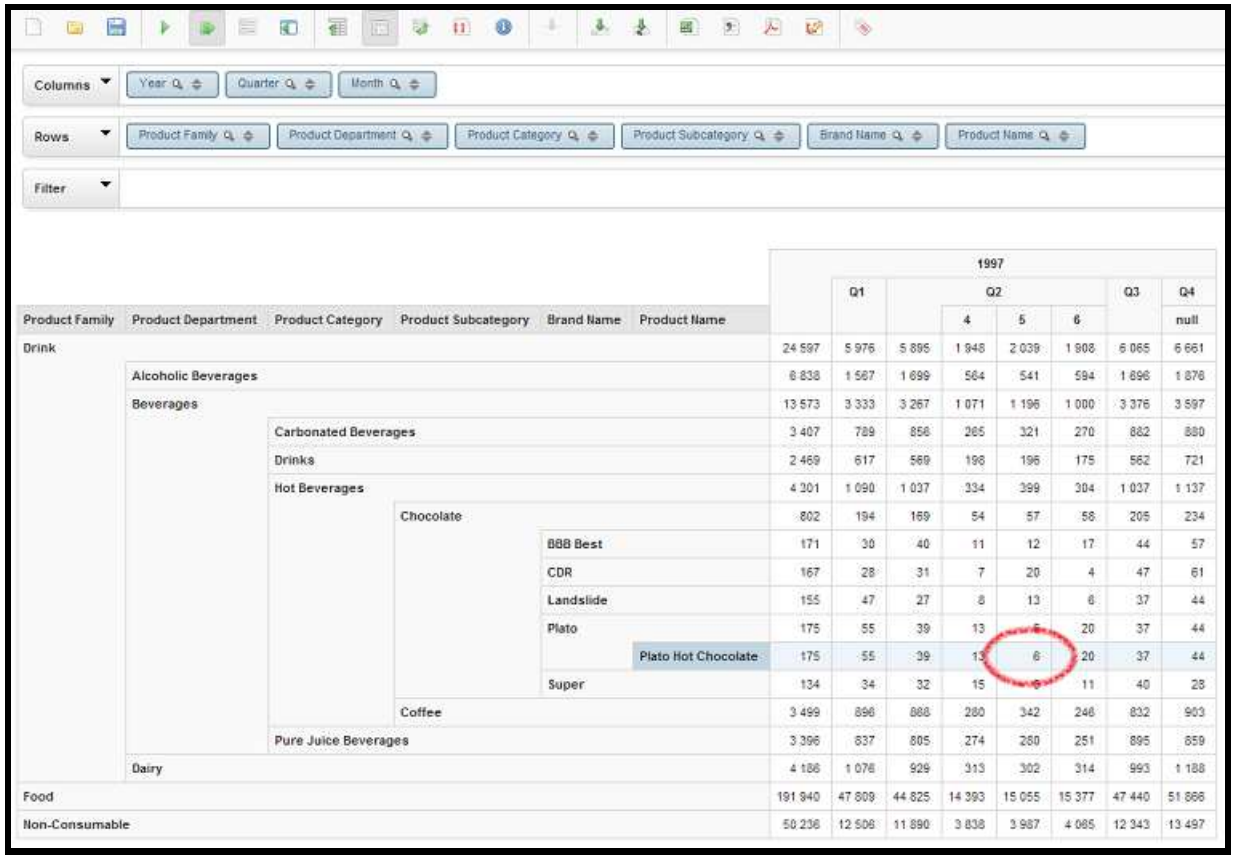


# Exporting - DEMO




| Media Type | $10K - $30K | $110K - $130K | $130K - $150K | $150K + | $30K - $50K | $50K - $70K | $70K - $90K | $90K - $110K |
|---|---|---|---|---|---|---|---|---|
| Bulk Mail | 1 070 | 133 | 226 | 54 | 1 270 | 917 | 478 | 172 |
| Cash Register Handout | 1 598 | 272 | 291 | 125 | 2 229 | 1 031 | 930 | 221 |
| Daily Paper | 1 724 | 355 | 676 | 212 | 2 352 | 1 257 | 948 | 214 |
| Daily Paper, Radio | 1 636 | 186 | 353 | 58 | 2 339 | 1 211 | 819 | 289 |
| Daily Paper, Radio, TV | 2 289 | 309 | 458 | 258 | 3 052 | 1 781 | 1 011 | 355 |
| In-Store Coupon | 843 | 184 | 363 | 136 | 1 191 | 554 | 406 | 121 |
| No Media | 41 607 | 8 496 | 10 167 | 4 088 | 64 456 | 32 891 | 24 518 | 9 225 |
| Product Attachment | 1 470 | 303 | 410 | 139 | 2 511 | 1 456 | 924 | 331 |
| Radio | 594 | 157 | 74 | 77 | 614 | 483 | 372 | 83 |
| Street Handout | 1 179 | 262 | 409 | 147 | 1 811 | 984 | 769 | 192 |
| Sunday Paper | 1 147 | 167 | 212 | 49 | 1 438 | 646 | 472 | 208 |
| Sunday Paper, Radio | 1 327 | 361 | 393 | 83 | 1 884 | 919 | 751 | 247 |
| Sunday Paper, Radio, TV | 548 | 236 | 196 | 67 | 1 027 | 302 | 253 | 97 |
| TV | 918 | 140 | 164 | 156 | 1 136 | 535 | 394 | 164 |

# Drill down - DEMO



Columns: Year | Quarter | Month

Rows: Product Family | Product Department | Product Category | Product Subcategory | Brand Name | Product Name

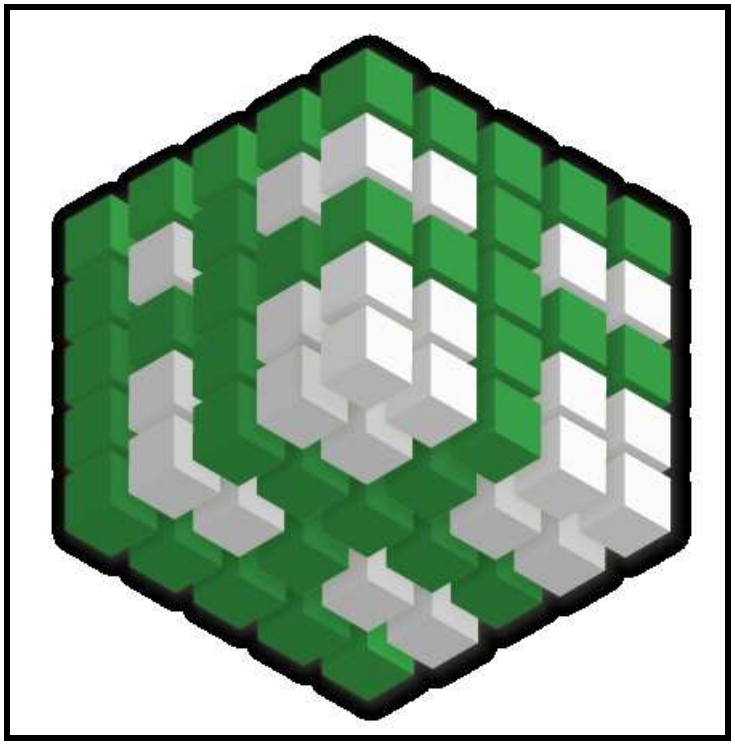
Filter:

| | | | | | | 1997 | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | Q1 | Q2 | | | | Q3 | Q4 |
| Product Family | Product Department | Product Category | Product Subcategory | Brand Name | Product Name | | | | 4 | 5 | 6 | | null |
| Drink | | | | | | 24 597 | 5 976 | 5 895 | 1 948 | 2 039 | 1 908 | 6 065 | 6 661 |
| | Alcoholic Beverages | | | | | 6 838 | 1 567 | 1 699 | 564 | 541 | 594 | 1 896 | 1 876 |
| | Beverages | | | | | 13 573 | 3 333 | 3 267 | 1 071 | 1 196 | 1 000 | 3 376 | 3 597 |
| | | Carbonated Beverages | | | | 3 407 | 789 | 856 | 265 | 321 | 270 | 882 | 880 |
| | | Drinks | | | | 2 469 | 617 | 569 | 196 | 196 | 175 | 562 | 721 |
| | | Hot Beverages | | | | 4 201 | 1 090 | 1 037 | 334 | 399 | 304 | 1 037 | 1 137 |
| | | | Chocolate | | | 802 | 194 | 169 | 54 | 57 | 56 | 205 | 234 |
| | | | | BBB Best | | 171 | 30 | 40 | 11 | 12 | 17 | 44 | 57 |
| | | | | CDR | | 167 | 28 | 31 | 7 | 20 | 4 | 47 | 61 |
| | | | | Landslide | | 155 | 47 | 27 | 8 | 13 | 6 | 37 | 44 |
| | | | | Plato | | 175 | 55 | 39 | 13 | 6 | 20 | 37 | 44 |
| | | | | | Plato Hot Chocolate | 175 | 55 | 39 | 13 | 6 | 20 | 37 | 44 |
| | | | | Super | | 134 | 34 | 32 | 15 | 6 | 11 | 40 | 28 |
| | | | Coffee | | | 3 499 | 896 | 868 | 280 | 342 | 246 | 832 | 903 |
| | | Pure Juice Beverages | | | | 3 396 | 837 | 805 | 274 | 280 | 251 | 895 | 859 |
| | Dairy | | | | | 4 186 | 1 076 | 929 | 313 | 302 | 314 | 993 | 1 188 |
| Food | | | | | | 191 940 | 47 809 | 44 825 | 14 393 | 15 055 | 15 377 | 47 440 | 51 866 |
| Non-Consumable | | | | | | 50 236 | 12 506 | 11 890 | 3 838 | 3 987 | 4 065 | 12 343 | 13 497 |

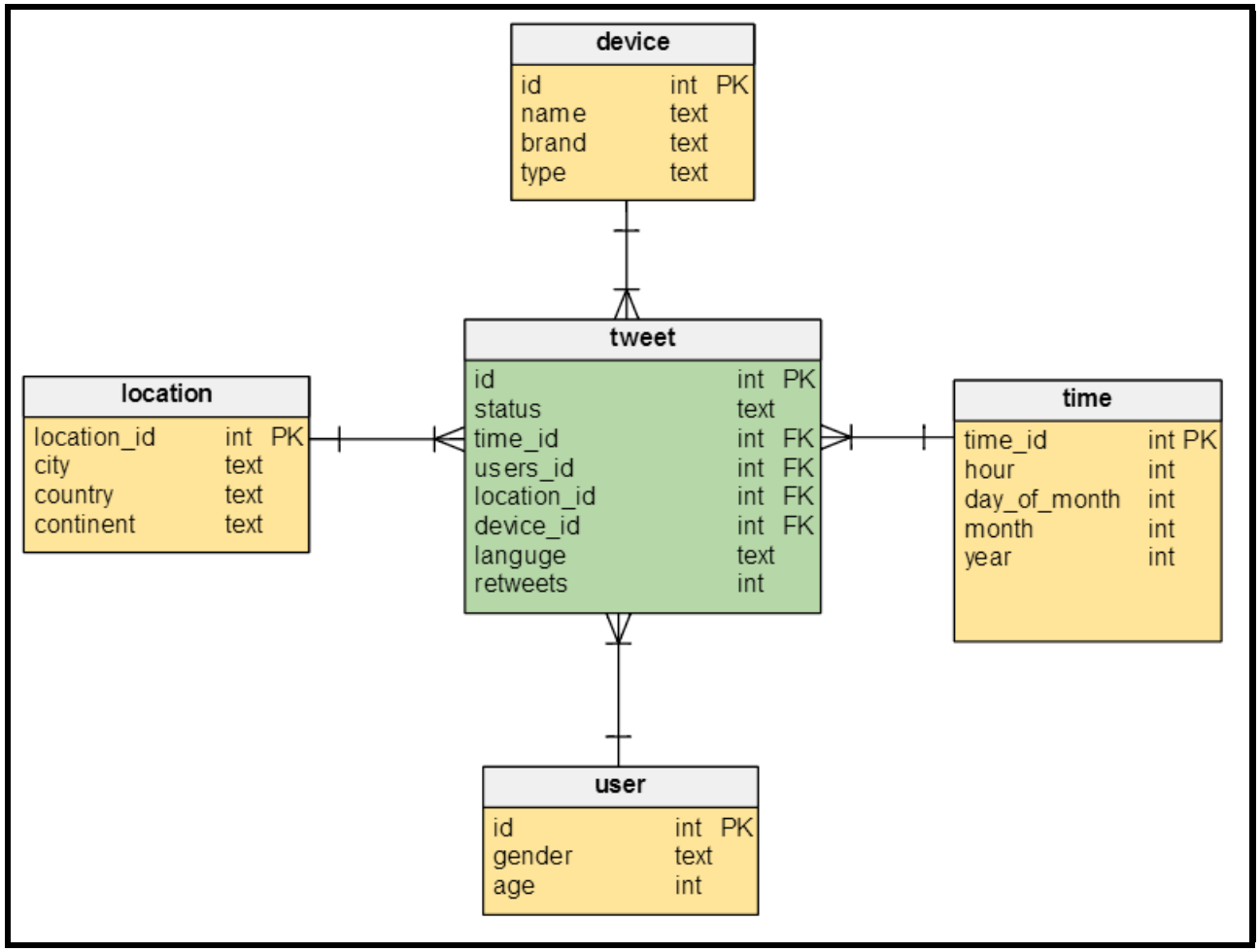
# Ignored concepts

- Hypercube
- Mondrian
- MDX

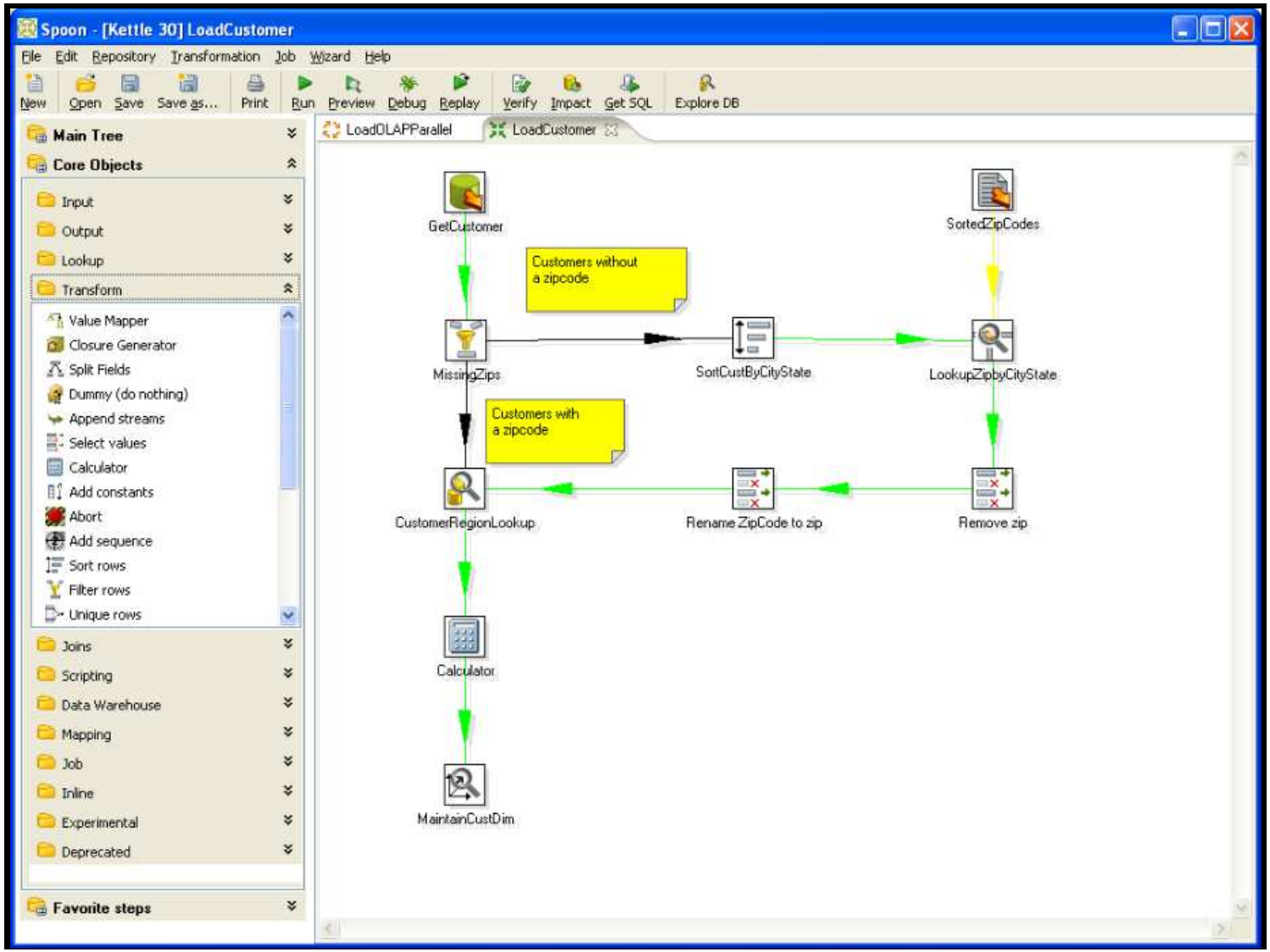
# Your own cube



# Star schema


device
id int PK
name text
brand text
type text
tweet
id int PK
status text
time_id int FK
users_id int FK
location_id int FK
device_id int FK
languge text
retweets int
location
location_id int PK
city text
country text
continent text
time
time_id int PK
hour int
day_of_month int
month int
year int
user
id int PK
gender text
age int

# ETL


Spoon - [Kettle 30] LoadCustomer
File Edit Repository Transformation Job Wizard Help
New Open Save Save as... Print Run Preview Debug Replay Verify Impact Get SQL Explore DB
Main Tree
Core Objects
Input
Output
Lookup
Transform
Value Mapper
Closure Generator
Split Fields
Dummy (do nothing)
Append streams
Select values
Calculator
Add constants
Abort
Add sequence
Sort rows
Filter rows
Unique rows
Joins
Scripting
Data Warehouse
Mapping
Job
Inline
Experimental
Deprecated
Favorite steps
LoadOLAPParallel
LoadCustomer
GetCustomer
SortedZipCodes
Customers without a zipcode
MissingZips
SortCustByCityState
LookupZipbyCityState
Customers with a zipcode
CustomerRegionLookup
Rename ZipCode to zip
Remove zip
Calculator
MaintainCustDim

# ETL - challenges

- Missing or incomplete data
- Heuristics
- Incremental, periodic updates
- Various data sources

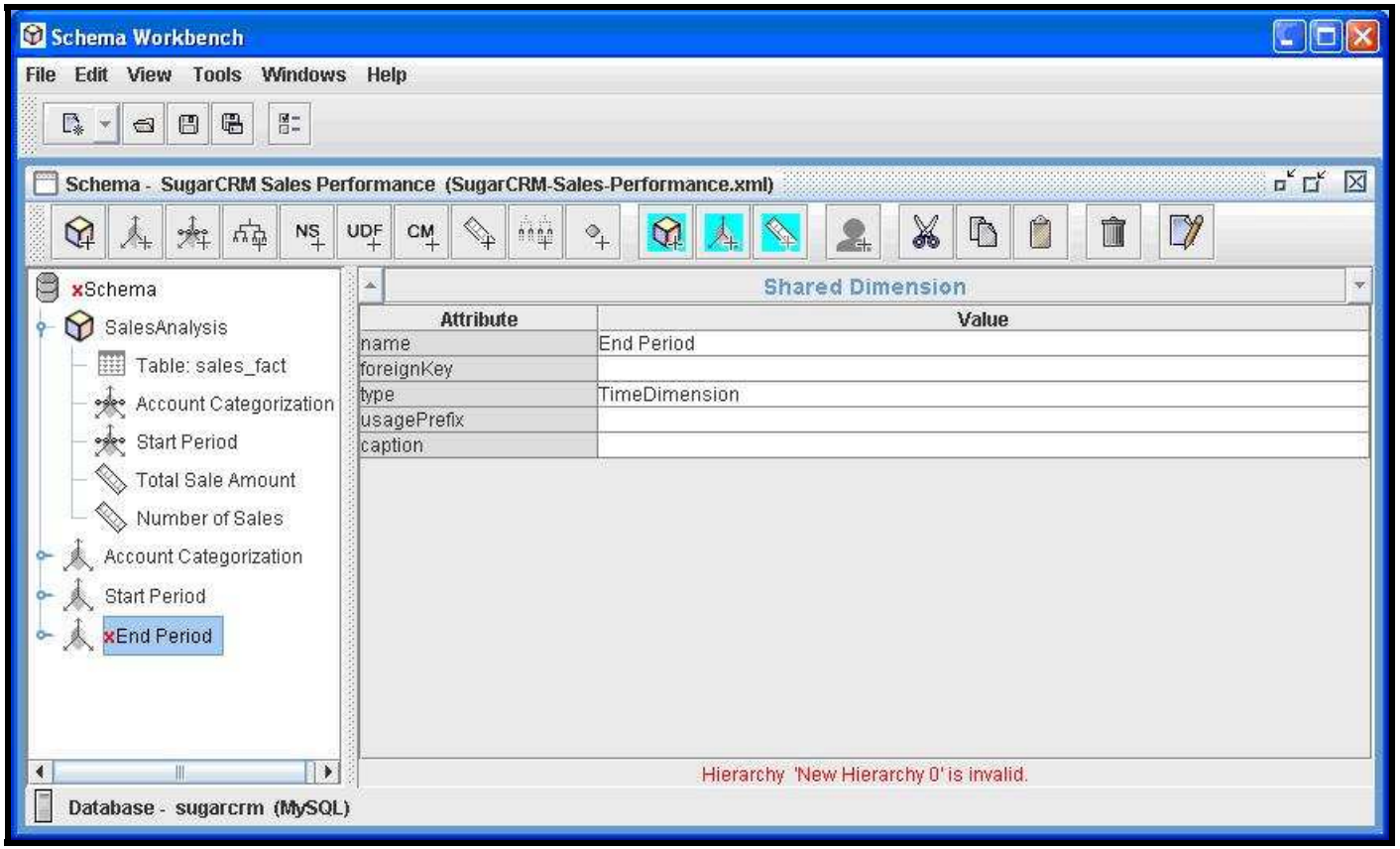
# Schema file

```
<Schema name="Twitter">

  <Cube name="Tweets" defaultMeasure="Count">
    <Table name="tweet">
    <DimensionUsage name="Time" source="Time"
foreignKey="time_id"/>
    <Dimension name="Location" foreignKey="location_id">
      <Hierarchy hasAll="true" allMemberName="All
locations">
        <Table name="location"/>
        <Level name="Continent" column="continent"/>
        <Level name="Country" column="country"/>
        <Level name="City" column="city"/>
      </Hierarchy>
    </Dimension>
    <!-- ... -->

</Schema>
```

# Schema Workbench




Source: www.stratebi.com/cursos/olap-mdx

# Security - users

- Standard user/password
- Roles
- Spring Security - customizable

# Security - data

- By role
- Restrict what can be seen
- Top/bottom limit

# Performance

Big data, *before it was cool*

- Indexes on foreign keys
- Aggregate tables

# Without Aggregate table

```
SELECT COUNT(id)
FROM tweet NATURAL JOIN locations
GROUP BY locations.continent
```

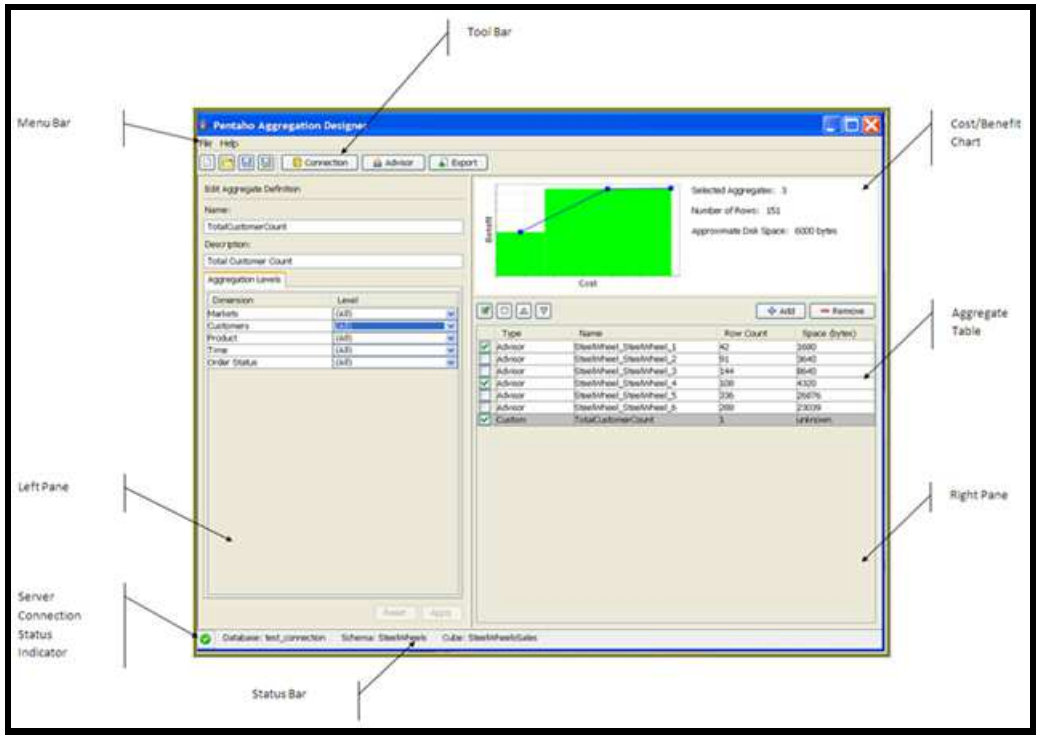
# With aggregate table

```
INSERT INTO agg (cnt, l.city, l.country, l.continent)
SELECT COUNT(t.id) AS cnt, city, country, continent
FROM tweet t NATURAL JOIN locations l
GROUP BY l.city
```

Usages:

```
SELECT SUM(agg.count)
FROM agg
GROUP BY locations.continent
```

# Pentaho Aggregation Designer




Source: infocenter.pentaho.com/help/index.jsp

# Deployment

- `mondrian.jar` - engine
- `saiku.war` - RESTful web services
- `ui.war` - JS front-end

# Disadvantages

- Horizontal scalability?
- Stuck with SQL databases
- Complex schema definition (XML)
- Aggregate tables are hard

# Thank you!

Slides: nurkiewicz.github.io/talks/2014/33degree