Concept*Mix*:
Self-Service Analytical Data Integration Based on the Concept-Oriented Model

Alexandr Savinov

Database Technology Group

Technische Universität Dresden, Germany

Data Commander - http://conceptoriented.com

ESF
Europäischer Sozialfonds für Deutschland

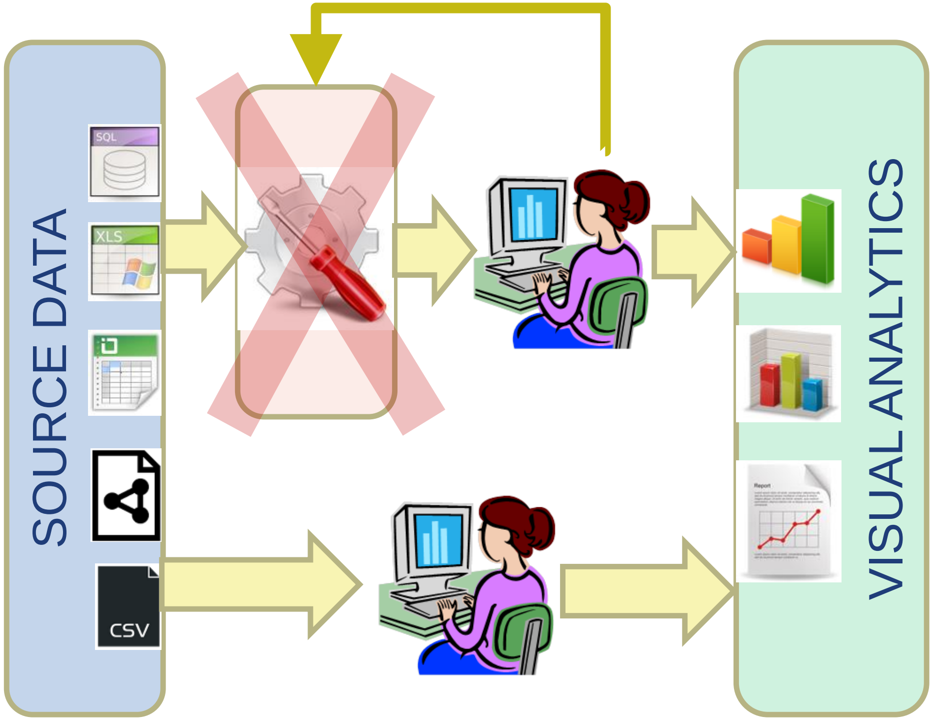
Gefördert durch:
Bundesministerium für Wirtschaft und Technologie
aufgrund eines Beschlusses des Deutschen Bundestages

eXIST
Existenzgründungen aus der Wissenschaft

dresden | exists
WISSEN. GRÜNDEN. UNTERNEHMEN.

PPROBLEM

- Variety of data sources: one aspect of the big data problem
  - Integrate: data sources have to be mashed up to produce the desired result
- Data wrangling (curation, munging, scraping) – the most tedious part of the overall analysis process
  - Transform: refactor the structure of data (schema)
- Original data does not have data the user needs
  - Analyze: new attributes have to be computed



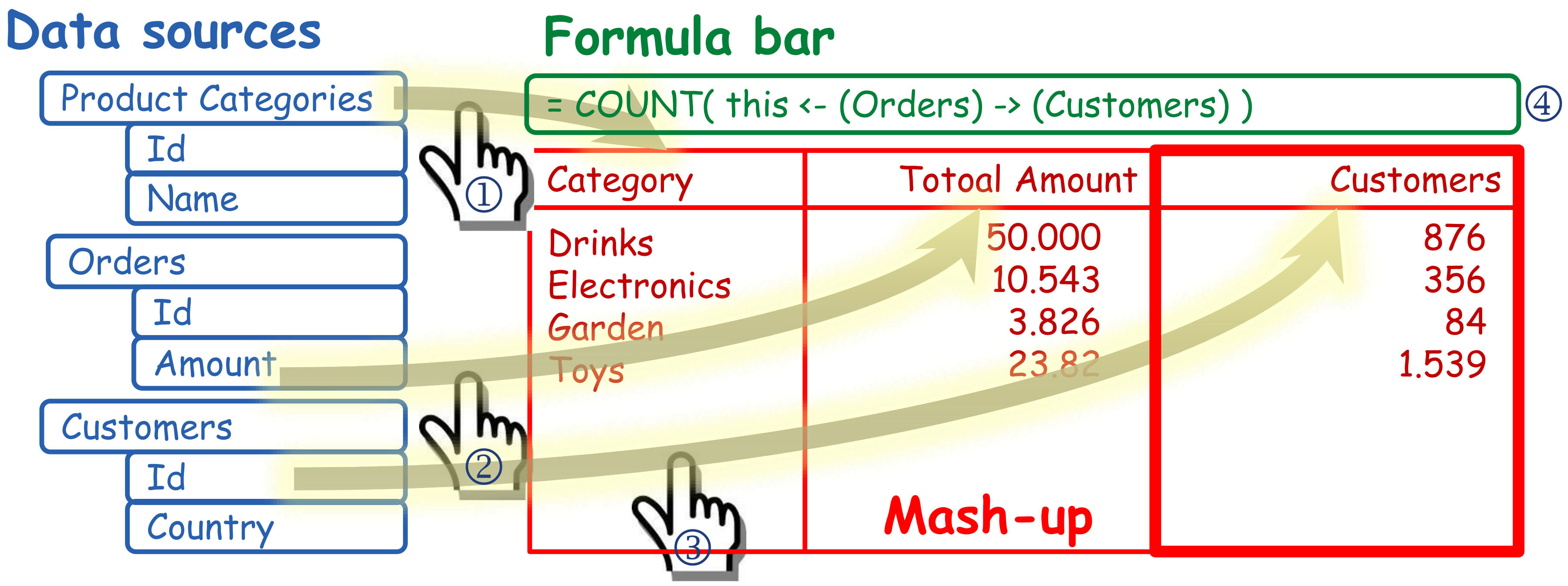
SELF-SERVICE & USER-DRIVEN

AD-HOC & AGILE

REAL-TIME & RESPONSIVE

Challenge: How to simplify operations with data so that the tool can be used by non-IT users?

PRODUCT VISION



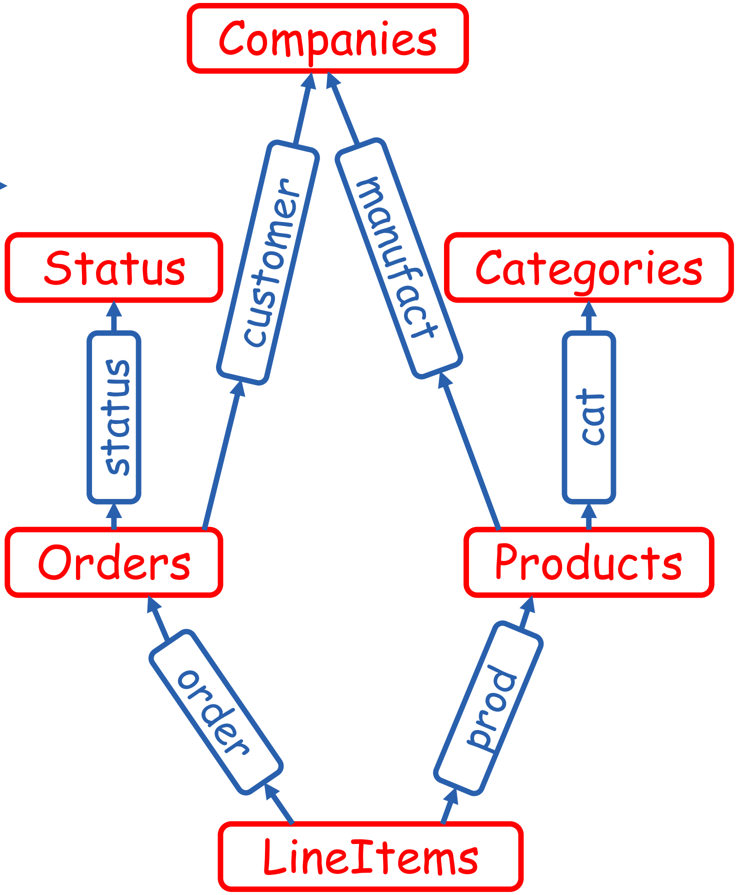
= COUNT(this <- (Orders) -> (Customers))

Category	Totoal Amount	Customers
Drinks	50.000	876
Electronics	10.543	356
Garden	3.826	84
Toys	23.82	1.539

- ConceptMix: self-service data integration, transformation and analysis tool
- Concept*Mix* is column-oriented rather than cell-oriented
- Data is defined by column formulas (4) rather than cell-formulas
- Drag-n-drop a source column (1-3) with automatic recommendations

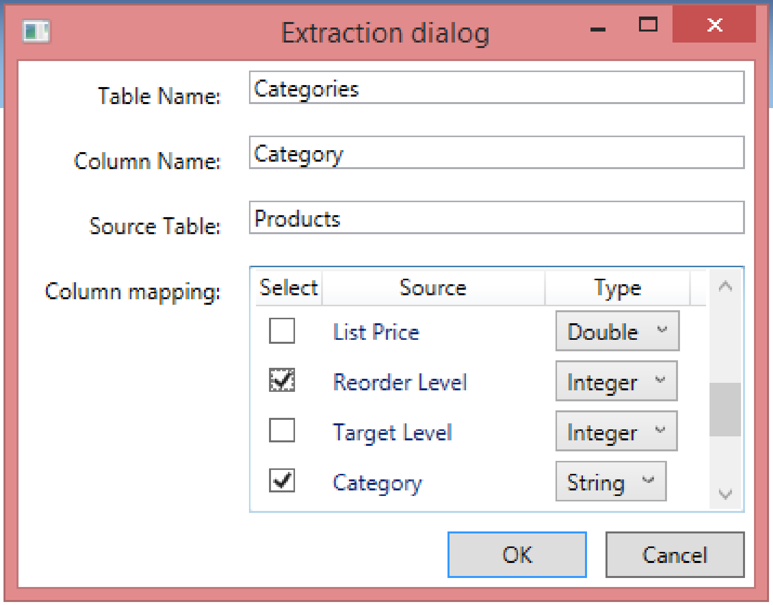
TECHNOLOGY

- Key enabler: concept-orientation:
  - Concept-oriented model of data (COM)
    - Unified model: simple and natural representation
    - Partially ordered set
    - Functional approach
  - Concept-oriented expression language (COEL)
    - No joins, no group-bys, no formal logic
    - Simple and expressive analytical operations
    - Algebra of functions
  - Column-based data processing model
    - Fast analytical operations with data (analytical database)
    - Column is a function
- More info: http://conceptoriented.org



SETS

- Goal: define a new set in terms of existing sets and functions
- Two operations
  - Product: all combinations of greater sets
  - Project: all outputs of some function



Extraction dialog

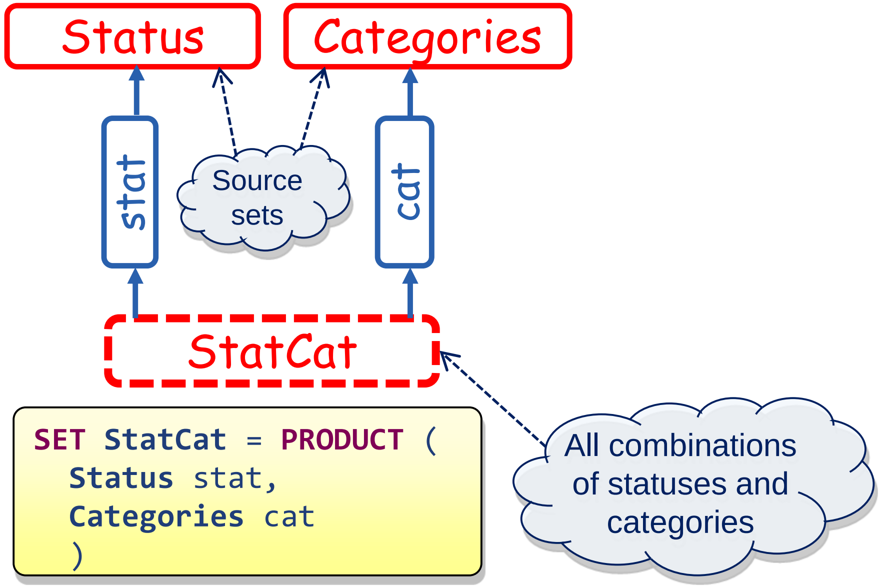
Table Name: Categories

Column Name: Category

Source Table: Products

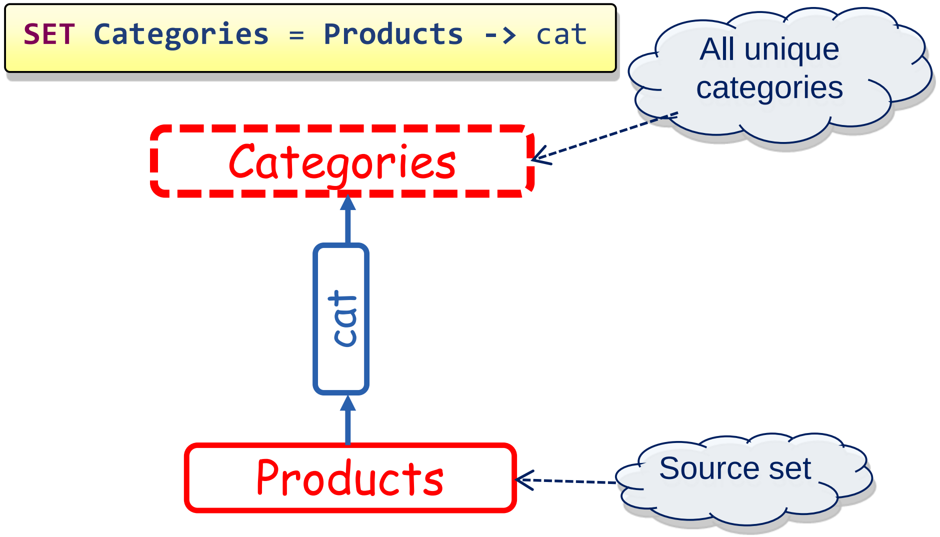
Column mapping:

Select	Source	Type
□	List Price	Double
☑	Reorder Level	Integer
□	Target Level	Integer
☑	Category	String



```
SET StatCat = PRODUCT (
  Status stat,
  Categories cat
 )
```

All combinations of statuses and categories



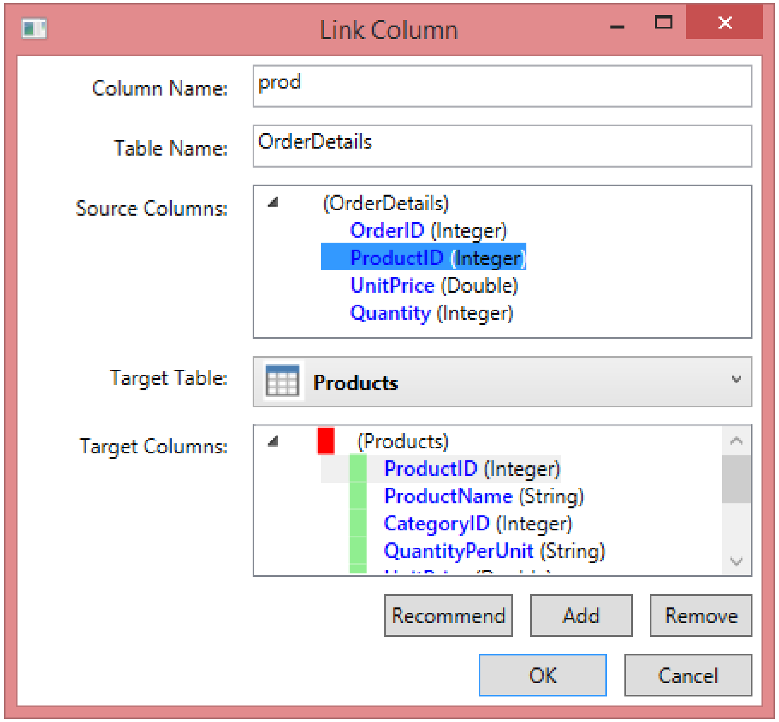
```
SET Categories = Products -> cat
```

All unique categories

Source set

LINKS

- Goal: link to sets using existing functions



Link Column

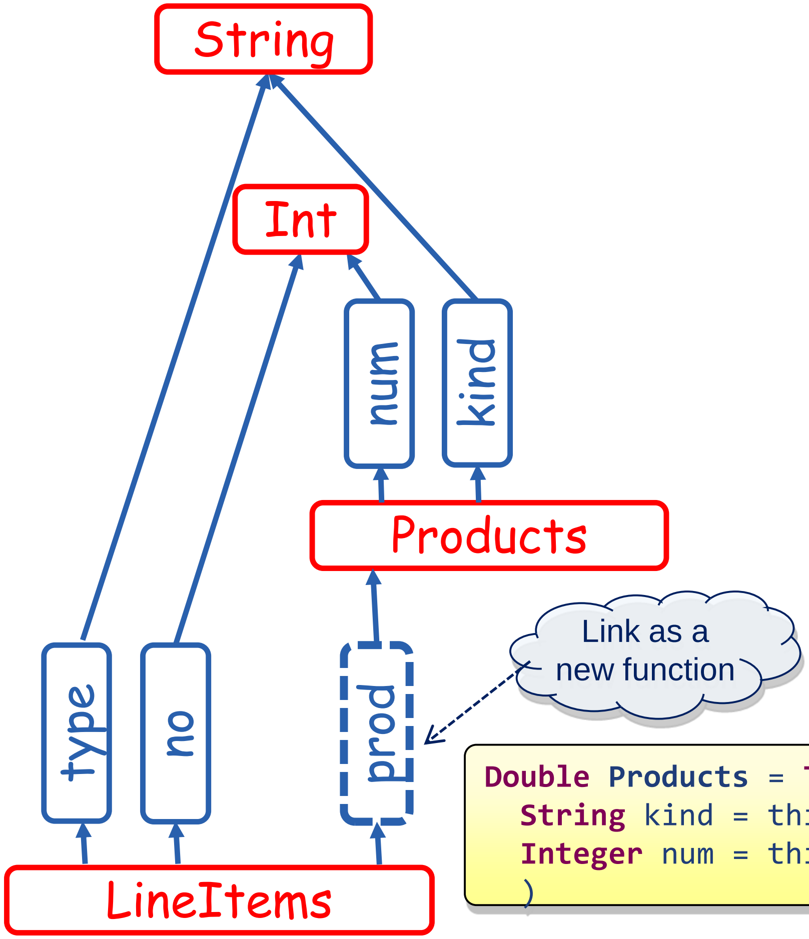
Column Name: prod

Table Name: OrderDetails

Source Columns:
- (OrderDetails)
  - OrderID (Integer)
  - ProductID (Integer)
  - UnitPrice (Double)
  - Quantity (Integer)

Target Table: Products

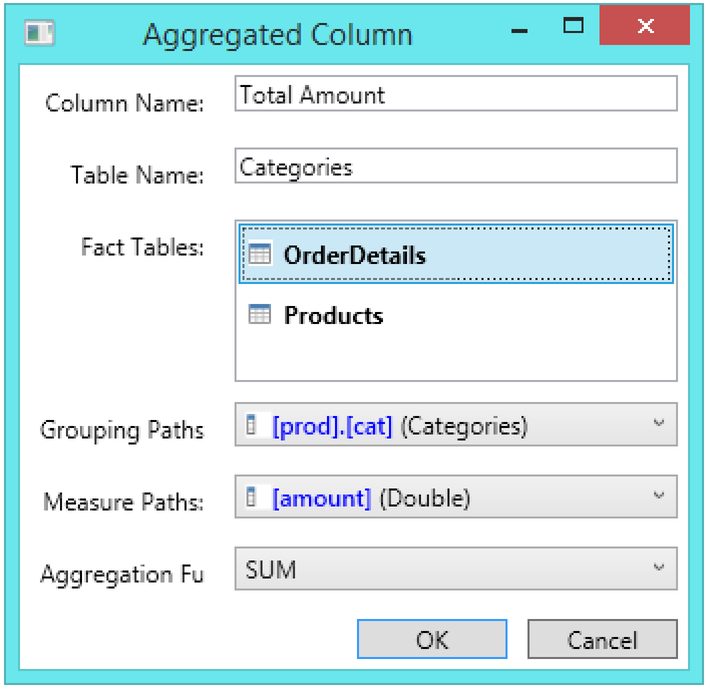
Target Columns:
- (Products)
  - ProductID (Integer)
  - ProductName (String)
  - CategoryID (Integer)
  - QuantityPerUnit (String)



```
Double Products = TUPLE (
  String kind = this.type,
  Integer num = this.no,
  )
```

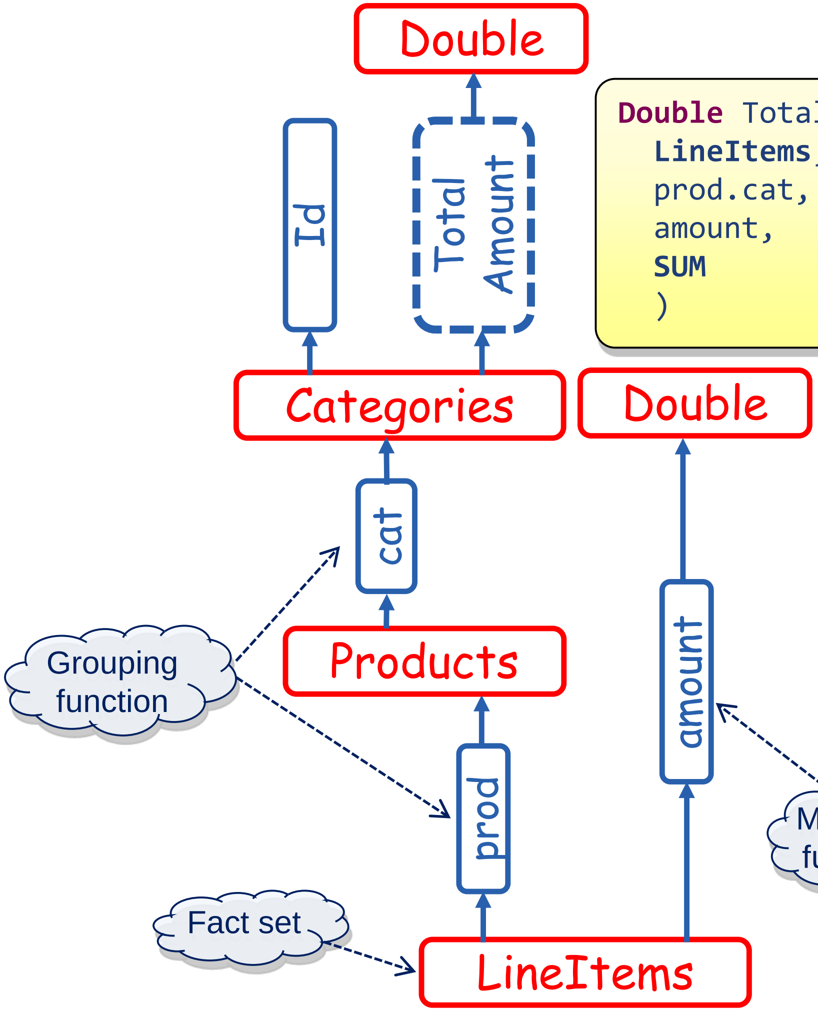
AGGREGATION

- Parameters:
  - Fact set
  - Grouping function
  - Measure function
  - Aggregation function



Aggregated Column

Column Name:	Total Amount
Table Name:	Categories
Fact Tables:	OrderDetails Products
Grouping Paths	[prod].[cat] (Categories)
Measure Paths:	[amount] (Double)
Aggregation Fu	SUM



```
Double TotalAmount = AGGREGATE (
  LineItems,
  prod.cat,
  amount,
  SUM
  )
```

CONCLUSION

- Novelties:
  - Unified data model and expression language are used
  - Column formulas as opposed to cell formulas for derived data
- Advantages of ConceptMix (Data Commander):
  - Ease of use: radically simplifies analytical data integration; kills complexities when manipulating data
  - Fast time-to-value: from months to minutes
  - Lower IT costs: move the burden of authoring BI contents to the end users
  - Increase motivation; more convenient consumption of data
- Future work:
  - Assistance engine: recommending mappings, relationships, sources
  - Selection propagation and inference for interactive analysis
- More info: http://conceptoriented.org