Multidimensional data representation and manipulation

Data Cube Concepts

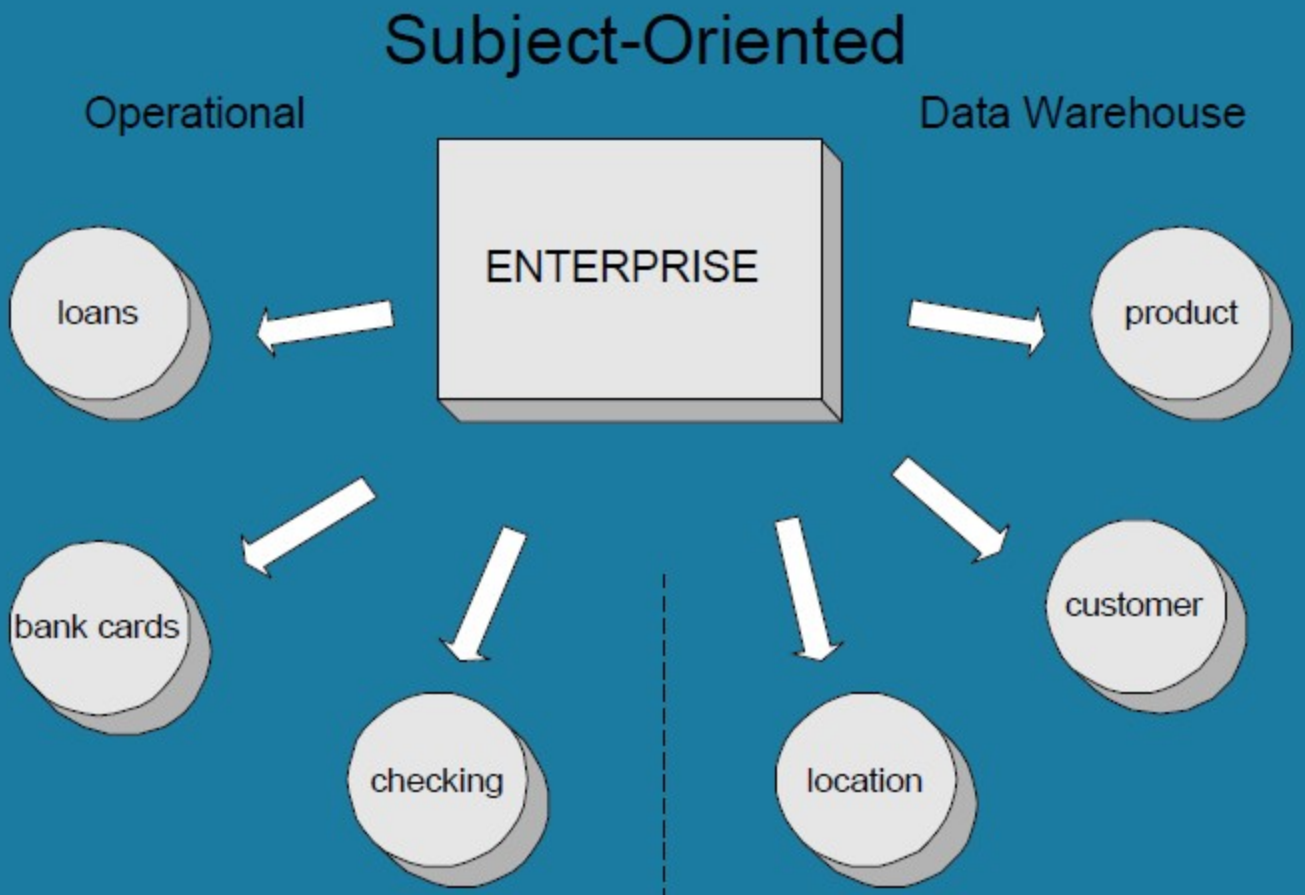
Data Warehouse Definitions

- Theoretical definition:

"A data warehouse is a subject-oriented, integrated, time-variant, nonvolatile collection of data in support of management's decision-making process"

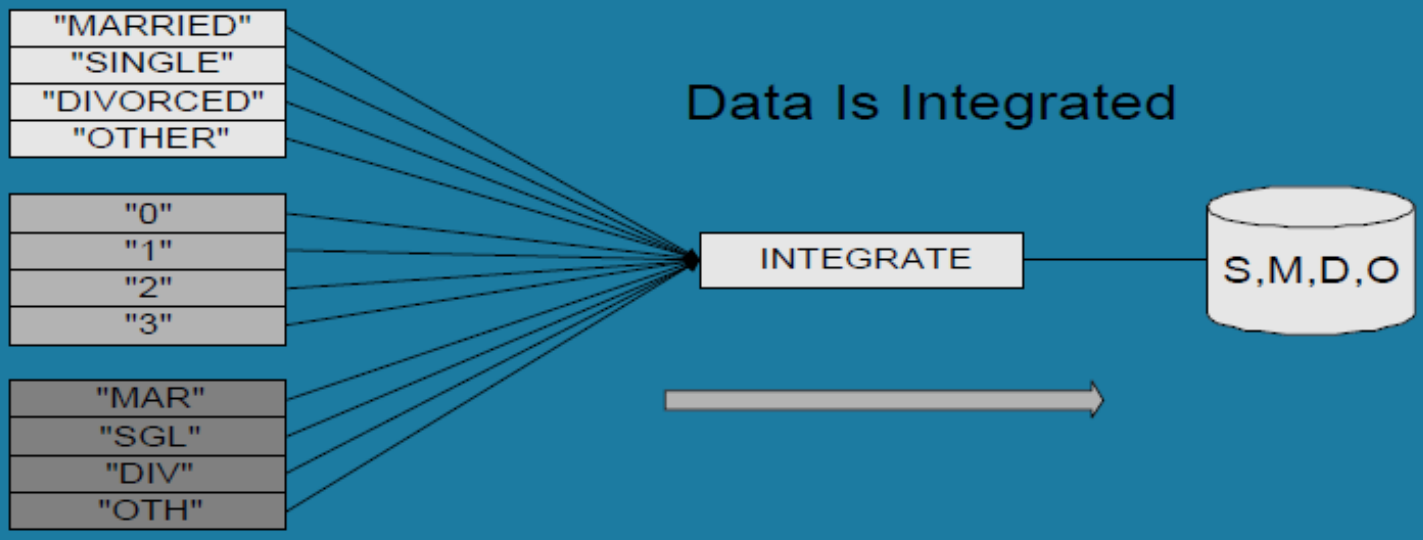
Using the Data Warehouse - Wiley, W.H. Inmon

Data Warehouse Definitions (cont..)



Data Warehouse Definitions (cont..)

- Integrated
  - Data can come from many different sources
  - Each source can look different than the other
  - Once it's in the DW, it should look the same



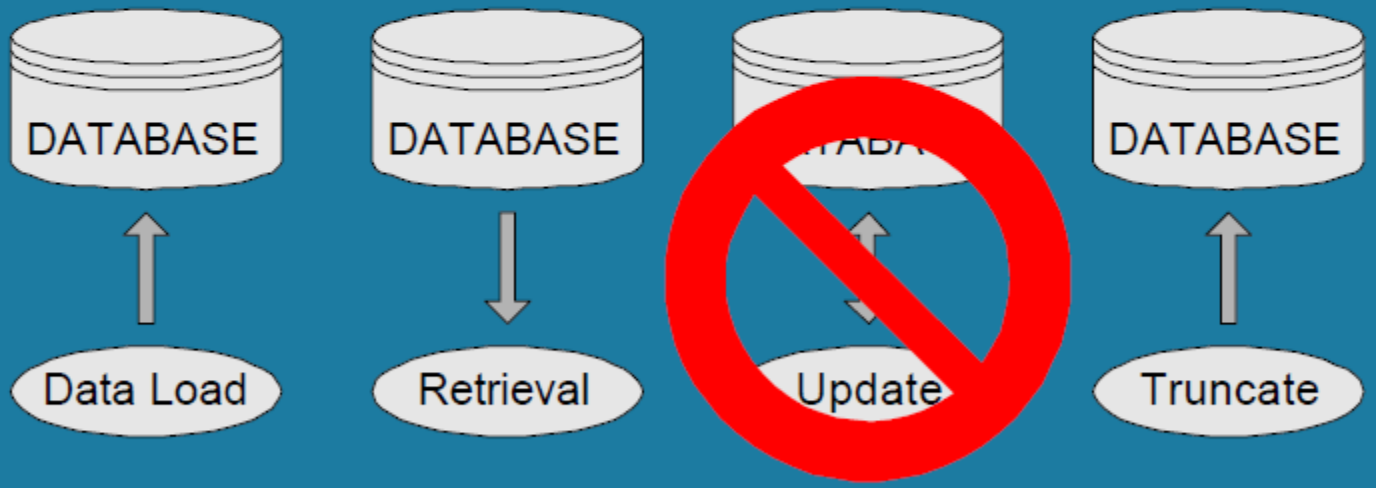
Data Warehouse Definitions (cont..)

- Time-Variant
  - Key structure is an element of time
  - No matter how it's organized, it still represents a series of snapshots
  - Snapshots or slices can lose accuracy over time, as opposed to the operational environment, which doesn't lose accuracy.

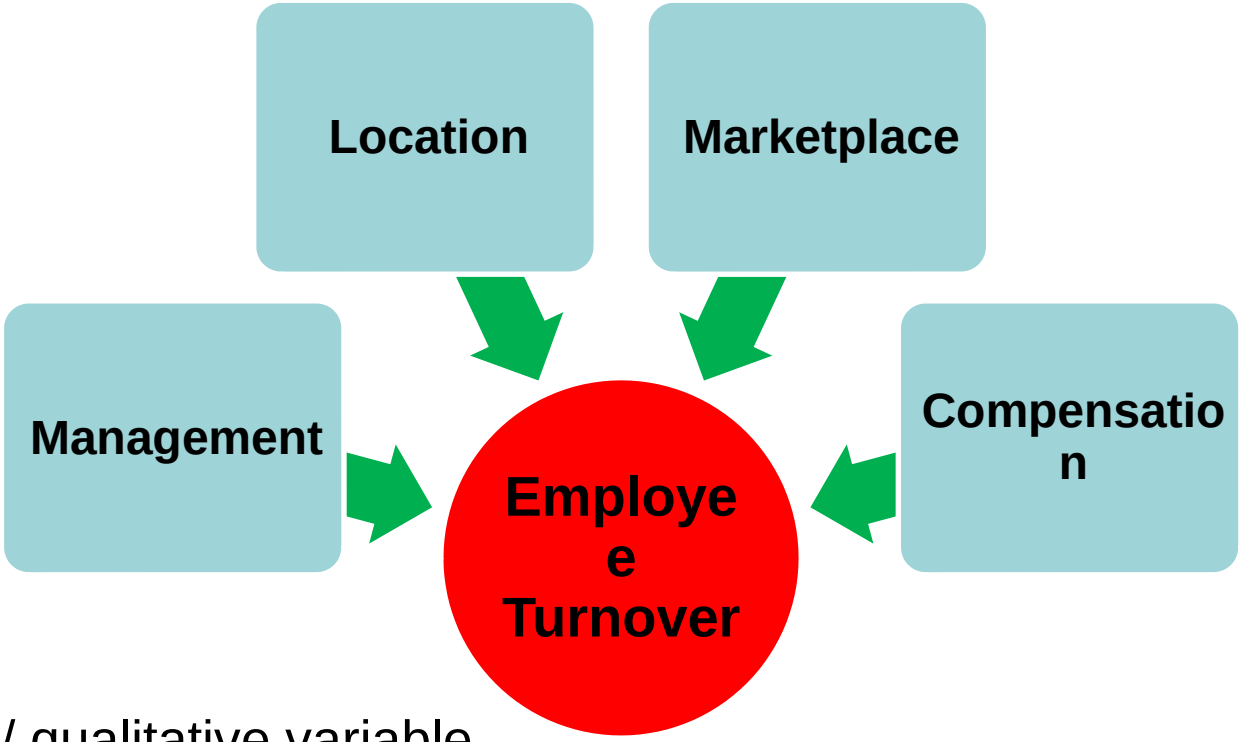
    Example: "Our product code has changed since last year."

Data Warehouse Definitions (cont..)

- Nonvolatile
  - Two types of routine operations: Load & Access
  - No update
  - Removal of data according to business rule



Business Analyst Perspective



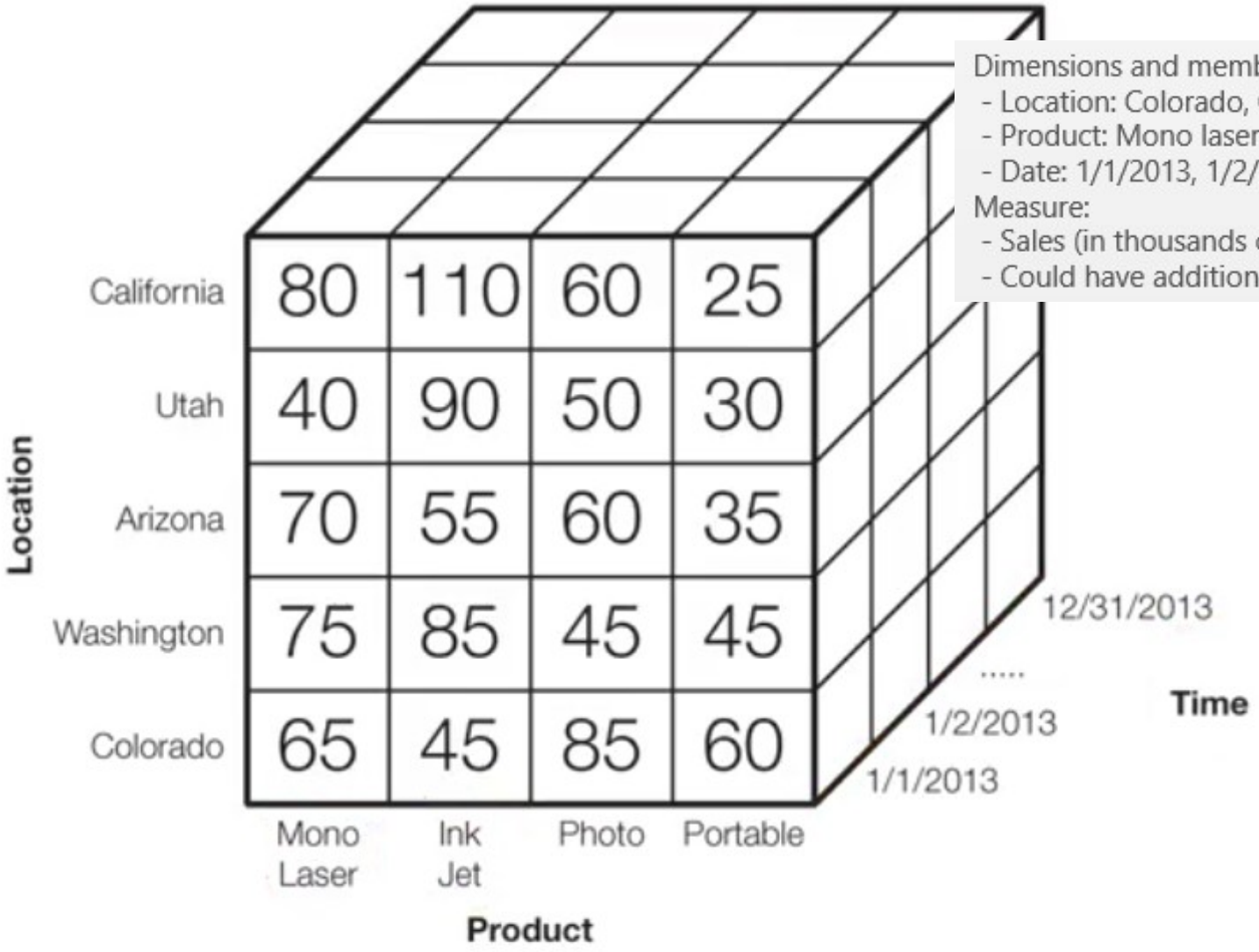
Factor / qualitative variable

Outcome Variable / quantitative variable

Data Cube Basics

- Business analyst model
  - Factors or influencing variables of interest
  - Quantitative variables
  - Multidimensional arrangement
- Terminology
  - Dimension: subject label for a row or column
    Can have more than 2 or 3 dimensions
    dimension may be city size or type of health plan offered
  - Member: value of dimension
  - Measure: quantitative variables/data stored in cells
    can have more than one measure in a cell

Sales Data Cube Example



Dimensions and members:
- Location: Colorado, California, ...
- Product: Mono laser, Ink Jet, ...
- Date: 1/1/2013, 1/2/2013

Measure:
- Sales (in thousands of dollars)
- Could have additional measures such as unit sales

Notes on Dimensions and Measures

- Hierarchies:
  - Member can have sub members (more detail)
  - Location: country, region, state, zip code
- Sparsity:
  - Many cells are typically empty when dimensions are related
  - May not sell all products in all regions
  - Major problem with storing data cubes:
    - compression of unused space
- Measures:
  - Derived measure:
    - Common: unit sales * unit volume; sales per transaction
    - Data cube engine must compute efficiently
  - Multiple measures in cells

- Granularity (level of details or summarization of the units of data)
  - high level of granularity contains low level of detail
  - low level of granularity contains high level of detail

- Sparsity (low density)
  - Data is normally stored in sparse form. If no value exists for a given combination of dimension values, no row exists in the fact table. For example, if not every product is sold in every market. In this case, Market and Product are sparse dimensions.

  - DENSE DATA: Most multidimensional databases may also contain dense dimensions. A fact table is considered to have dense data if it has (of a high probability to have) one row for every combination of its associated dimension levels.

Measure Aggregation Properties

- “Aggregate Property” indicates allowable summary operations for measures
- Additive
  - Summarized by addition across all dimensions such as sales, cost, profit
  - Sales can be summed across product, time, customer, …
- Semi-Additive
  - Summarized by addition in some but not all dimensions such as time
  - Periodic measurements such as account balances and inventory levels
  - Account balance can be summed across customer branch
  - Account balance cannot be summed across time because balance is just a point in time measurement
- Non-Additive
  - Cannot be summarized by addition through any dimension
  - Historical facts such as unit price for a sale

Measure Aggregation Example

- Dimensions
  - Course: course id, degree, department, and college
  - Student: student id, major, department, and college
  - Time: semester, academic year, academic decade
- Measures:
  - Credit hours
  - Grade
  - Unit tuition (cost per credit hour)
  - Tuition (unit tution * credit hours)
- Aggregation properties for measures: ?

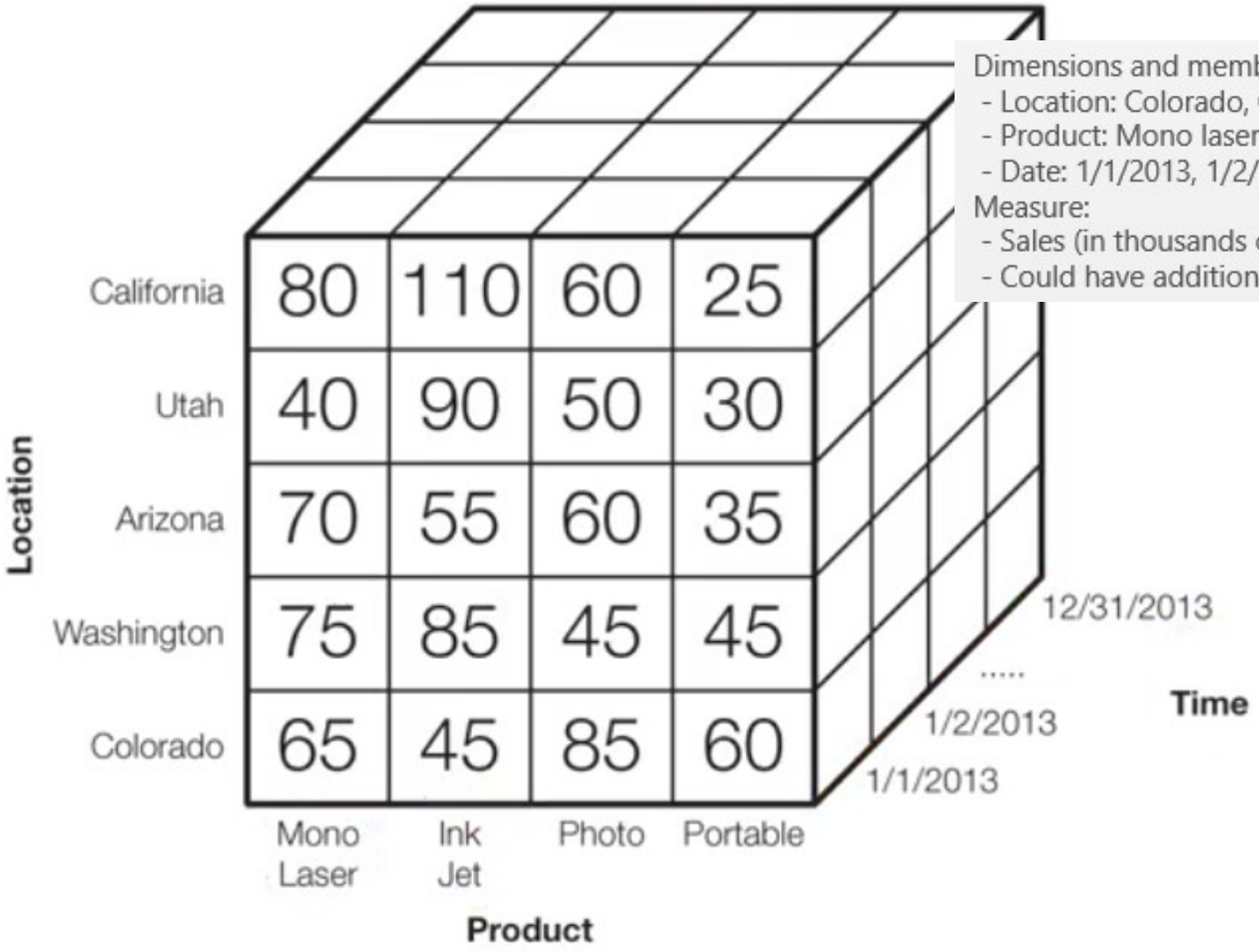
  Credit hours: additive across all dimensions

  Grade: non additive but averageable such as grade point average

  Unit tuition: non additive and averageable but probably not useful as an average

  Tuition: additive

Sales Data Cube Example



Dimensions and members:
- Location: Colorado, California, ...
- Product: Mono laser, Ink Jet, ...
- Date: 1/1/2013, 1/2/2013

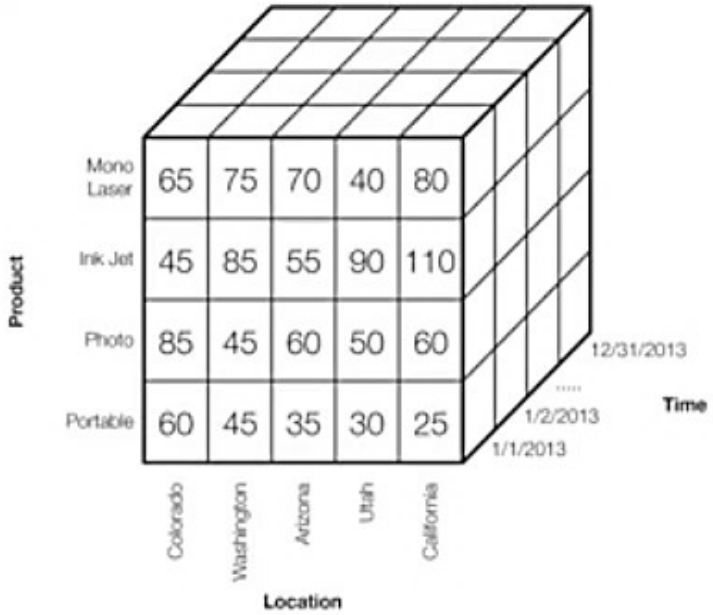
Measure:
- Sales (in thousands of dollars)
- Could have additional measures such as unit sales

WHAT IS THE GENERIC MEANING OF THE VERB “PIVOT”?

WHAT DOES PIVOT MEAN FOR A DATA CUBE?

SLICE OPERATOR

- Subset of dimensions
- Set dimension to specific value

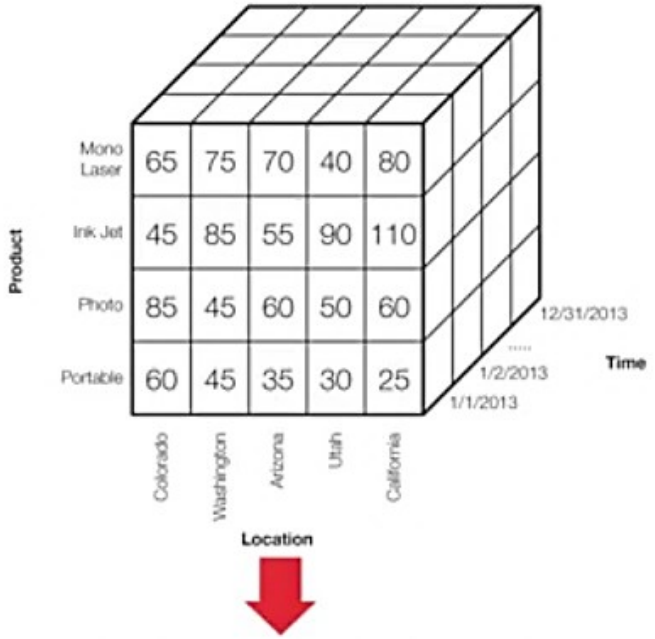


(Location x Product Slice for Time = 1/1/2013)

Location	Product			
	Mono Laser	Ink Jet	Photo	Portable
California	80	110	60	25
Utah	40	90	50	30
Arizona	70	55	60	35
Washington	75	85	45	45
Colorado	65	45	85	60

SLICE SUMMARIZE VARIATION

- Replace a dimension with a summary of its values across all members



(Location x Time Slice SUM Product Sales)

Location	Time			
	1/1/2013	1/2/2013	...	Total Sales
California	275	670	...	16,250
Utah	210	190	...	11,107
Arizona	220	255	...	21,500
Washington	250	285	...	20,900
Colorado	255	245	...	21,336

DICE OPERATOR

- Replace a dimension with a subset of values
- Dice operation often follows a slice operation

Location	Product			
	Mono Laser	Ink Jet	Photo	Portable
California	80	110	60	25
Utah	40	90	50	30
Arizona	70	55	60	35
Washington	75	85	45	45
Colorado	65	45	85	60

(Utah, Colorado, Arizona Dice)

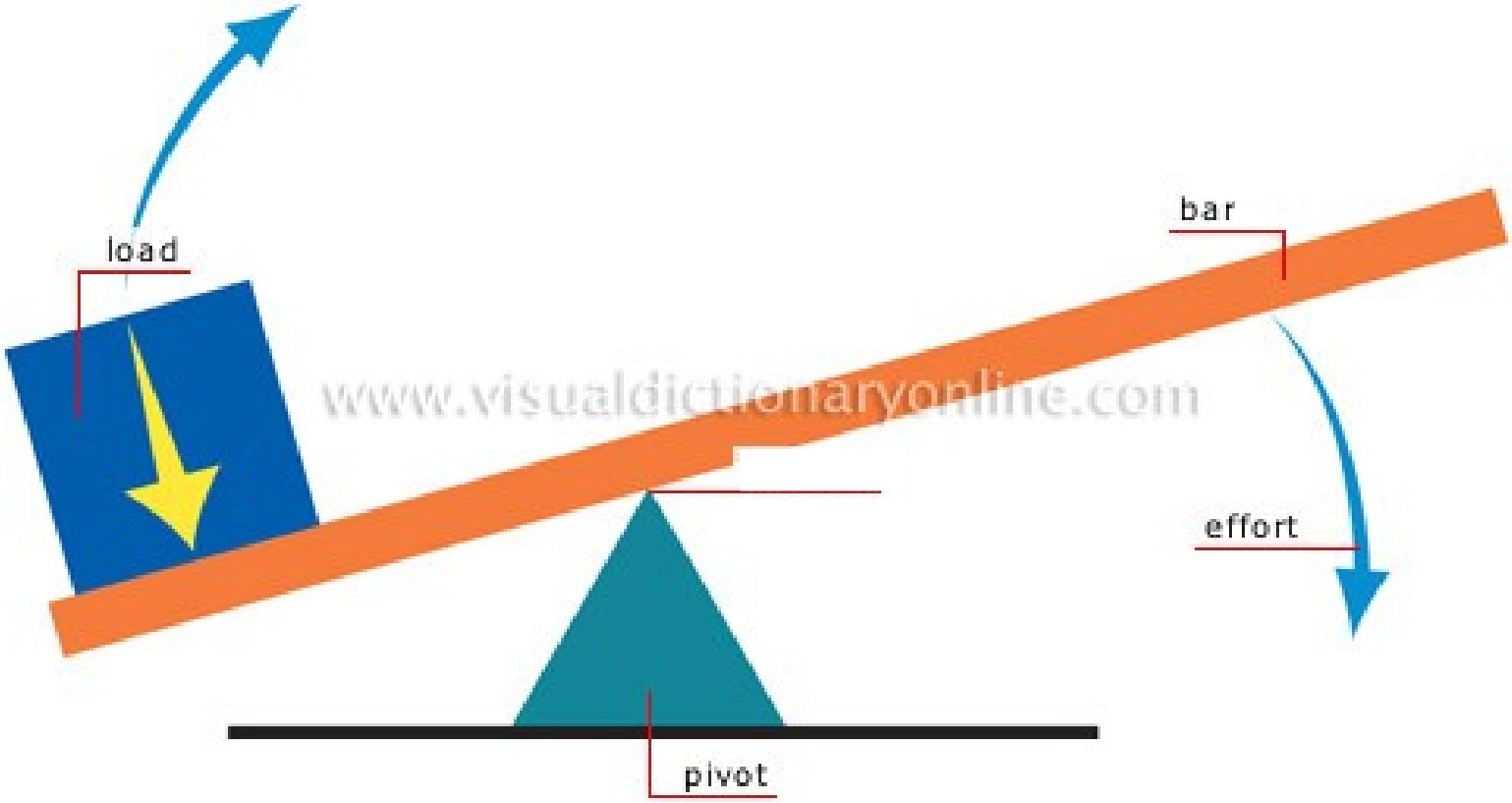
Location	Product			
	Mono Laser	Ink Jet	Photo	Portable
Utah	40	90	50	30
Arizona	70	55	60	35
Colorado	65	45	85	60

NAVIGATION OPERATORS

- Operators for hierarchical dimensions
- Drill-down: add detail to a dimension
- Roll-up: remove detail from a dimension
- Distribute or recalculate measure values

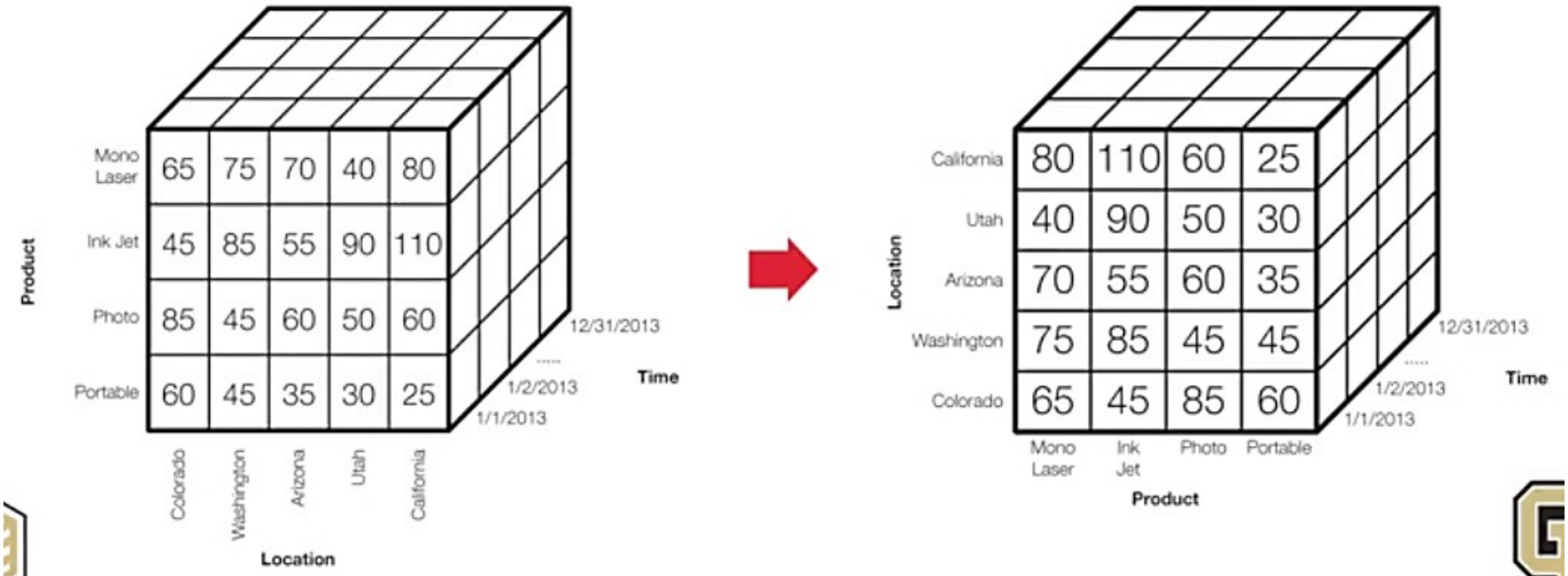
DRILL-DOWN EXAMPLE

Location	Product			
	Mono Laser	Ink Jet	Photo	Portable
California	80	110	60	25
−Utah				
Salt Lake	20	20	10	15
Park City	5	30	10	5
Ogden	15	40	30	10
Arizona	70	55	60	35
Washington	75	85	45	45
Colorado	65	45	85	60


load
bar
effort
pivot
www.visualdictionaryonline.com

PIVOT OPERATOR

- Rotate or rearrange dimensions

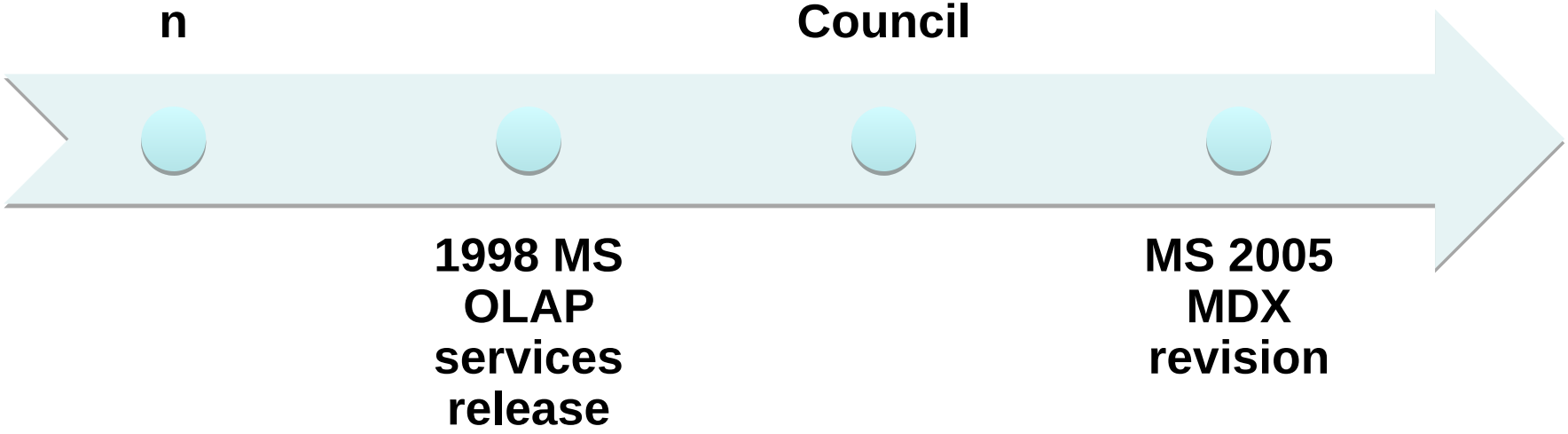


OPERATOR SUMMARY

Operator	Purpose	Description
Slice	Focus attention on a subset of dimensions	Replace a dimension with a single member value or with a summary of its measure values
Dice	Focus attention on a subset of member values	Replace a dimension with a subset of members
Drill-down	Obtain more detail about a dimension	Navigate from a more general level to a more specific level
Roll-up	Summarize details about a dimension	Navigate from a more specific level to a more general level
Pivot	Present data in a different order	Rearrange the dimensions in a data cube

Microsoft Multidimensional Expressions (MDX) Language History

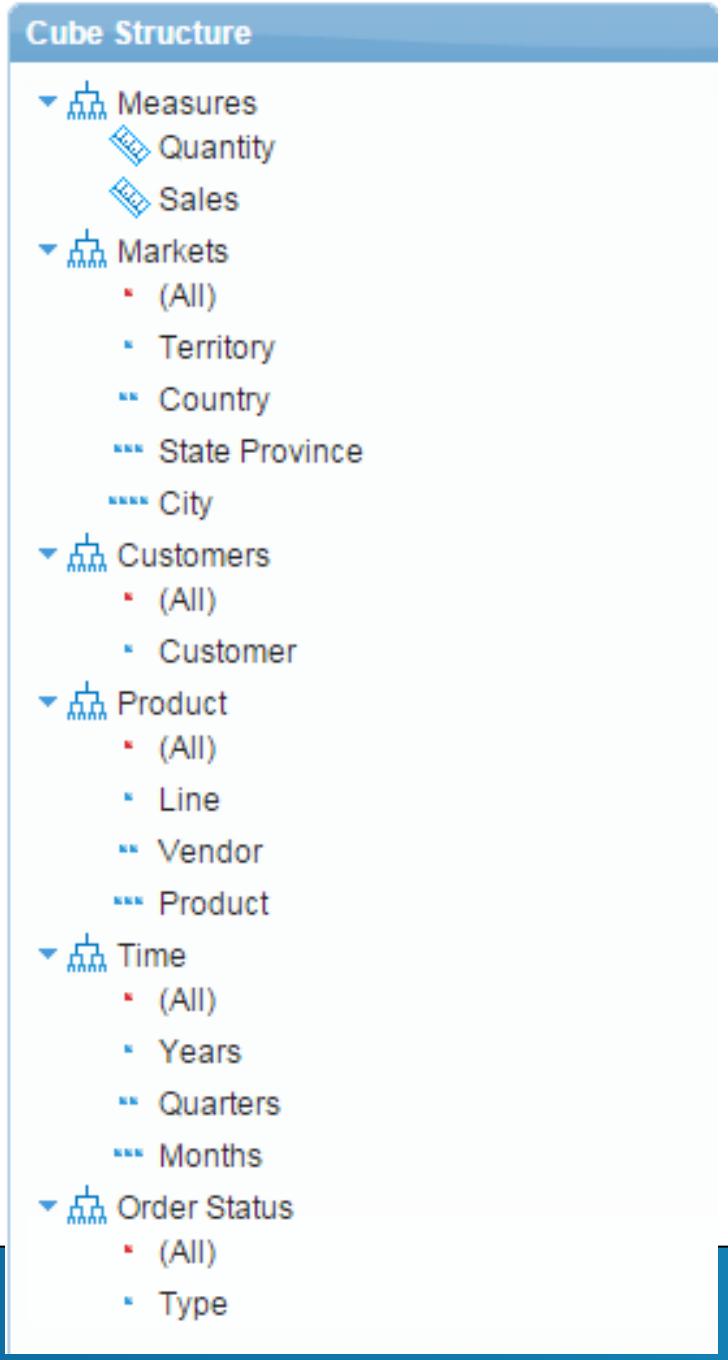
- Defacto standard developed by Microsoft and later by the XMLA (XML for Analysis) Council – A web standards group



MDX Usage

- Foundation for Microsoft products and open source analytics software
- SQL Server Analysis Services and Excel Pivot Tables
- Hyperion, IBM, SAP (Systems Applications and Products), and other vendors
- Foundation for open source projects: JPivot, Pivot4J, and Pentaho Business Analytics Platform

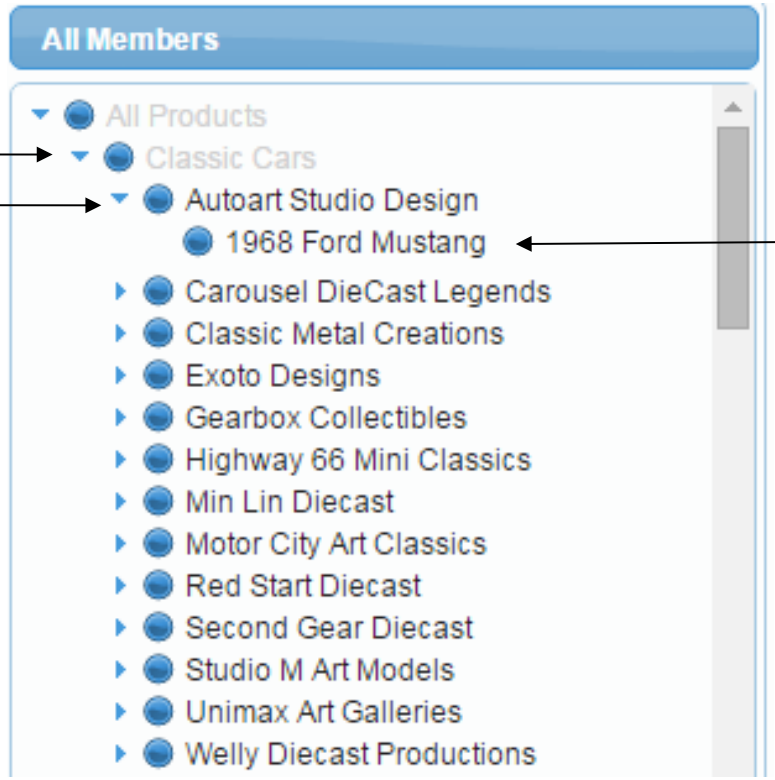
Example MDX Cube Structure



Attribute Hierarchy and Members

Line attribute member

Vendor attribute member



Product attribute member

Product attribute hierarchy

Steel Wheels Cube Display

	Time				
	All Years	2003	2004	2005	Average
	Measures	Measures	Measures	Measures	Measures
Product	Sales	Sales	Sales	Sales	Sales
Classic Cars	4,091,420	1,514,407	1,838,275	738,738	1,363,807
Motorcycles	1,274,125	397,220	590,580	286,325	424,708
Planes	1,076,757	347,755	528,928	200,074	358,919
Ships	748,671	244,821	375,672	128,178	249,557
Autoart Studio Design	67,592	19,764	36,027	11,801	22,531
Carousel DieCast Legends	208,583	75,184	102,537	30,862	69,528
Min Lin Diecast	79,662	29,691	37,098	12,873	26,554
Red Start Diecast	77,872	25,207	40,948	11,717	25,957
Studio M Art Models	84,190	27,795	39,390	17,005	28,063
Unimax Art Galleries	147,078	42,313	73,966	30,799	49,026
Welly Diecast Productions	83,693	24,867	45,706	13,120	27,898
Trains	234,469	72,802	124,750	36,917	78,156
Trucks and Buses	1,154,281	420,430	531,976	201,875	384,760
Vintage Cars	2,066,226	679,949	997,560	388,718	688,742
Average	818,919	282,876	383,672	152,371	272,973

MDX Terminology Notes

- Tuple
  - Cell identifier
  - One member from each dimension
- Axis: dimension selected in a query (source cube cells)
- Slicer: combination of dimension members (result cube cells)

SQL Versus MDX

- Table result for SQL SELECT statement
- Data cube result for MDX SELECT statement
- Different mathematical approaches for manipulating tables (e.g. relational algebra) and data cubes (e.g. matrix algebra)

Comparison of Clauses

Language		
Clause	**SQL**	**MDX**
SELECT	List of columns	List of axis dimensions (source cube cells)
FROM	List of tables	Cube name
WHERE	Conditions restricting rows	Restriction to a combination of dimension members (result cube cells)

Example MDX Statement and Result

Query Result

Filter | Product

	Measures	
Time	Sales	Quantity
2003	1,514,407	12,762
2004	1,838,275	16,085

MDX Query

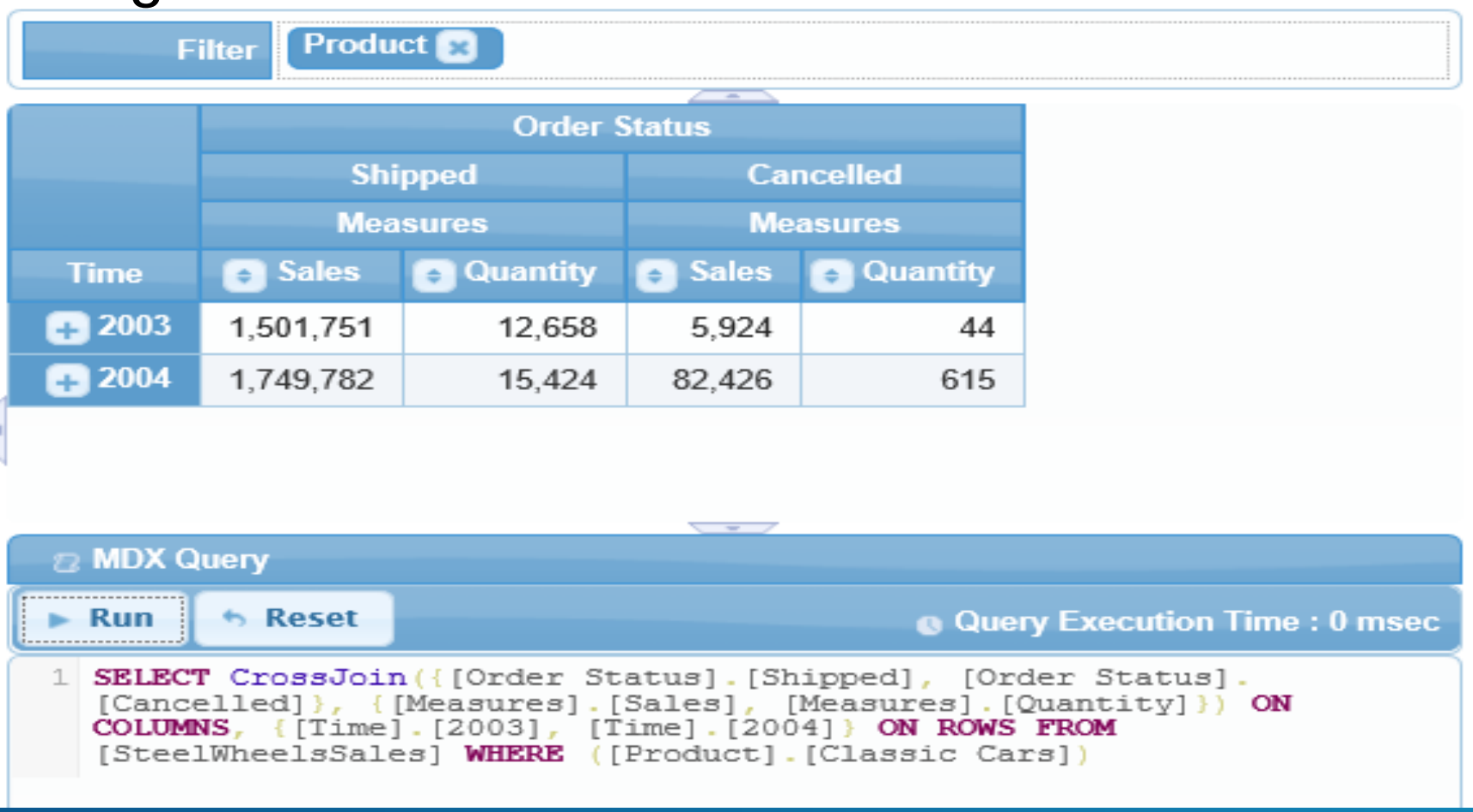
Run | Reset

```
1 SELECT {[Measures].[Sales], [Measures].[Quantity]} ON COLUMNS,
2 {[Time].[2003], [Time].[2004]} ON ROWS
3 FROM [SteelWheelsSales]
4 WHERE ([Product].[Classic Cars])
```

- **Dimensions in the WHERE clause must be different than the SELECT clause**
- **WHERE condition is known as a slicer condition**

CrossJoint Operation

- Combines multiple dimensions or measures on a single axis



Slicer Comparison Examples

	Order Status		
	All Status Types		
	Time		
Product	2003	2004	2005
Classic Cars	12,762	16,085	6,705
Motorcycles	4,031	5,906	2,771
Planes	3,833	5,820	2,207
Ships	2,844	4,309	1,346
Trains	1,000	1,409	409
Trucks and Buses	4,056	5,024	1,921
Vintage Cars	7,913	10,864	4,116

MDX Query

Run | Reset | Query Execution Time : 6 msec

```
1 SELECT CrossJoin({[Order Status].[All Status Types]}, {[Time].
[2003], [Time].[2004], [Time].[2005]}) ON COLUMNS, {[Product].
[Classic Cars], [Product].[Motorcycles], [Product].[Planes],
[Product].[Ships], [Product].[Trains], [Product].[Trucks and
Buses], [Product].[Vintage Cars]} ON ROWS FROM
[SteelWheelsSales]
```

	Order Status		
	All Status Types		
	Time		
Product	2003	2004	2005
Classic Cars	4,959	5,017	2,105
Motorcycles	1,744	2,809	568
Planes	977	2,224	592
Ships	702	1,642	537
Trains	409	326	177
Trucks and Buses	1,289	2,563	597
Vintage Cars	3,268	3,576	1,871

MDX Query

Run | Reset | Query Execution Time : 8 msec

```
1 SELECT CrossJoin({[Order Status].[All Status Types]}, {[Time].
[2003], [Time].[2004], [Time].[2005]}) ON COLUMNS, {[Product].
[Classic Cars], [Product].[Motorcycles], [Product].[Planes],
[Product].[Ships], [Product].[Trains], [Product].[Trucks and
Buses], [Product].[Vintage Cars]} ON ROWS FROM
[SteelWheelsSales] WHERE Markets.Territory.NA
```

Pivot Table

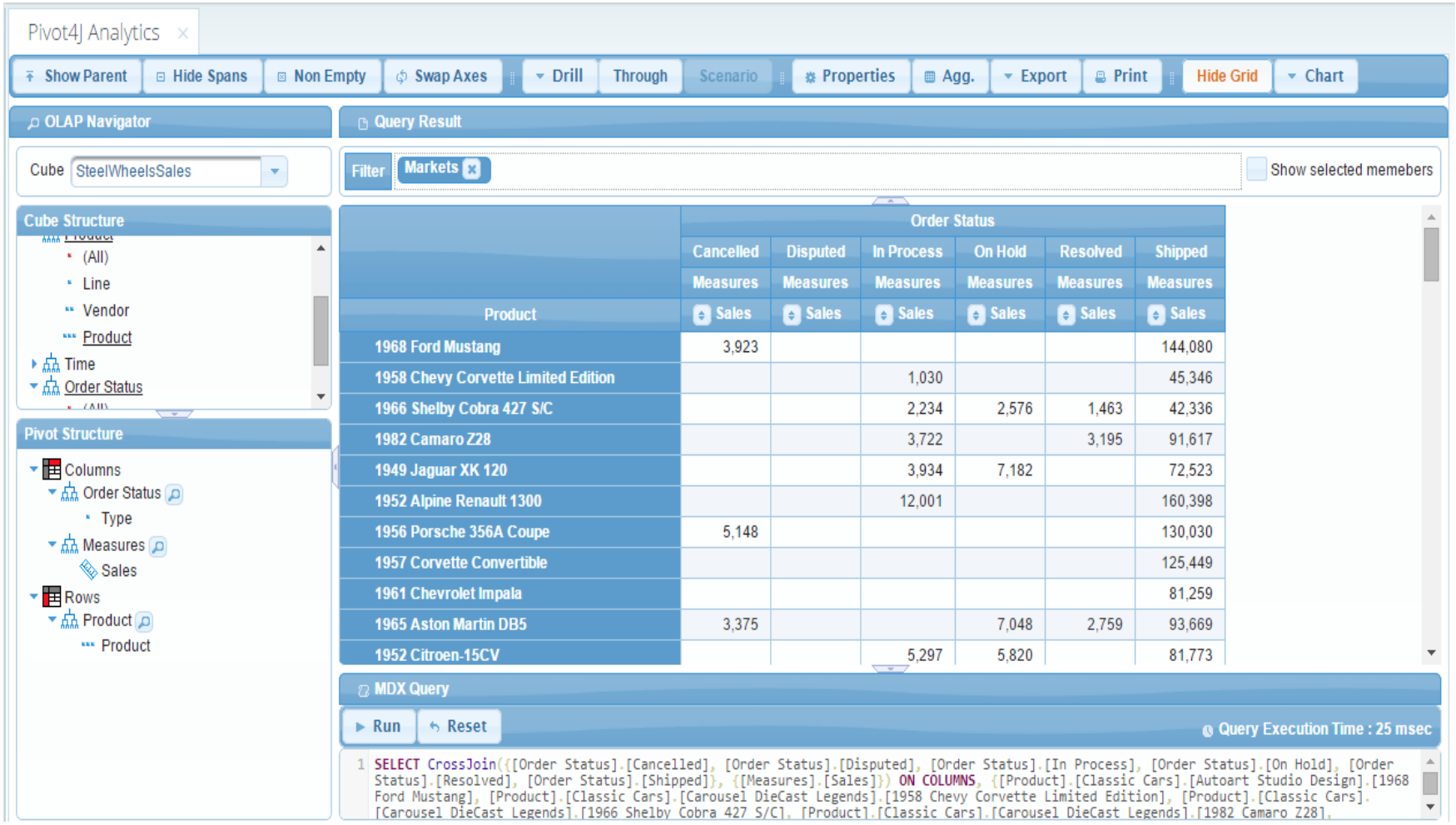
	Order Status						
	Cancelled	Disputed	In Process	On Hold	Resolved	Shipped	Total
	Measures	Measures	Measures	Measures	Measures	Measures	Measures
Product	Sales	Sales	Sales	Sales	Sales	Sales	Sales
1968 Ford Mustang	3,923					149,346	153,268
1958 Chevy Corvette Limited Edition			1,030			46,205	47,235
1966 Shelby Cobra 427 S/C			2,234	2,576	1,463	42,336	48,608
1982 Camaro Z28			3,722		3,195	97,362	104,280
1949 Jaguar XK 120			3,934	7,182		72,523	83,639
1952 Alpine Renault 1300			12,001			179,072	191,073
1956 Porsche 356A Coupe	5,148					135,479	140,627
1957 Corvette Convertible						137,115	137,115
1961 Chevrolet Impala						83,389	83,389
1965 Aston Martin DB5	3,375			7,048	2,759	93,669	106,851
1952 Citroen-15CV			5,297	5,820		81,773	92,890
1969 Chevrolet Camaro Z28			3,386	1,057		66,304	70,747
1992 Porsche Cayenne Turbo Silver	2,367					99,789	102,156
1948 Porsche 356-A Roadster	1,930					77,738	79,669

- Powerful interface for data cubes
- Convenient rearrangement of row and column headings
- Expand or collapse dimensions

Pivot4J

- Allows cube representation similar to pivot table in Microsoft Excel
- Works with Pentaho Business Analytics
- Separate add-on
- Graphical implementation of the MDX language

Pivot4J Interface



Pivot Table with MDX Statement

Query Result

Filter: Product

	Measures	
Time	Sales	Quantity
2003	1,514,407	12,762
2004	1,838,275	16,085

MDX Query

Run | Reset

```
1 SELECT {[Measures].[Sales], [Measures].[Quantity]} ON COLUMNS,
2 {[Time].[2003], [Time].[2004]} ON ROWS
3 FROM [SteelWheelsSales]
4 WHERE ([Product].[Classic Cars])
```

Pivot Table with CrossJoin