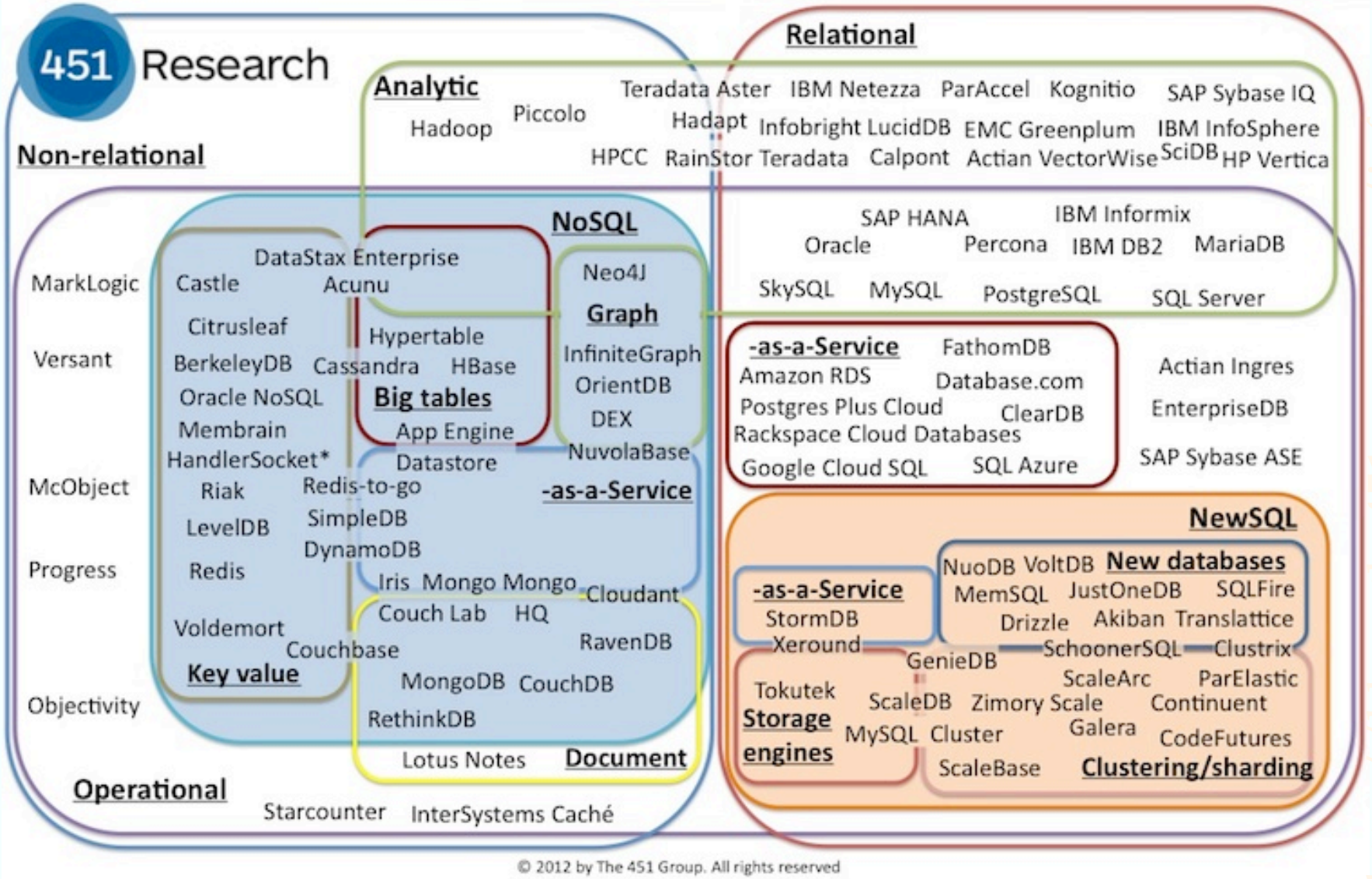
Data Warehousing and Decision Support

[R&G] Chapter 23, Part A

Machine Learning is not a business

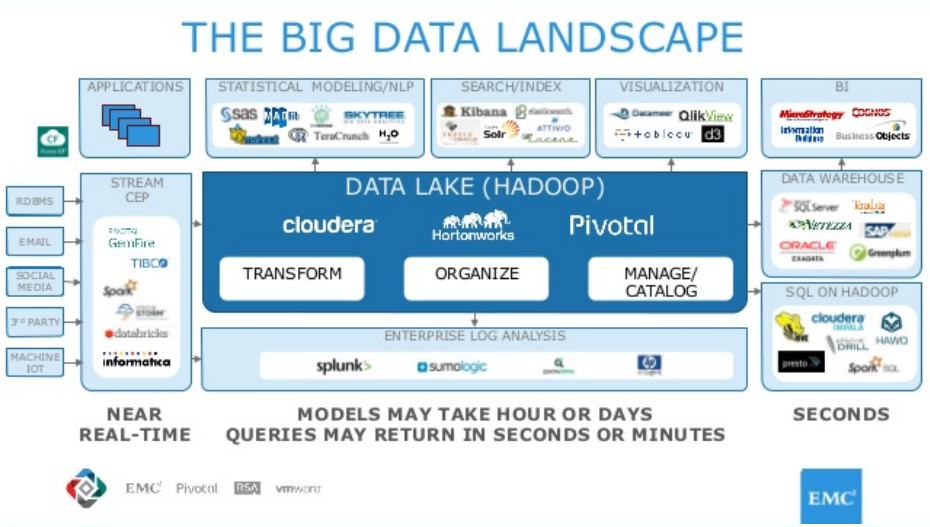
- Databases/data warehousing is!
- ~$50000/TB/year for DWH.
- Stable for decades.
- You can't sell a machine learning solution unless you package it with a data management system.
- But then, there's plenty of money to be made and nobody competes on price!

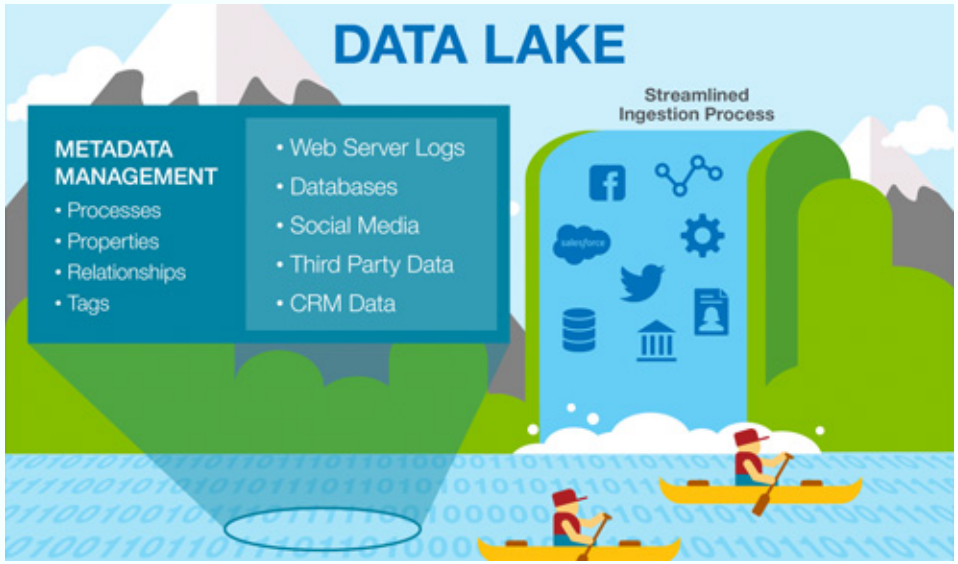
The evolving database landscape

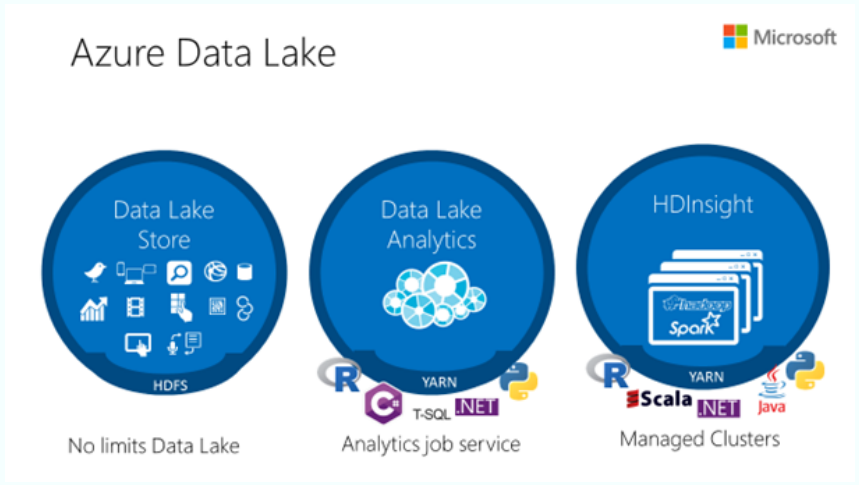


Data warehousing is big business

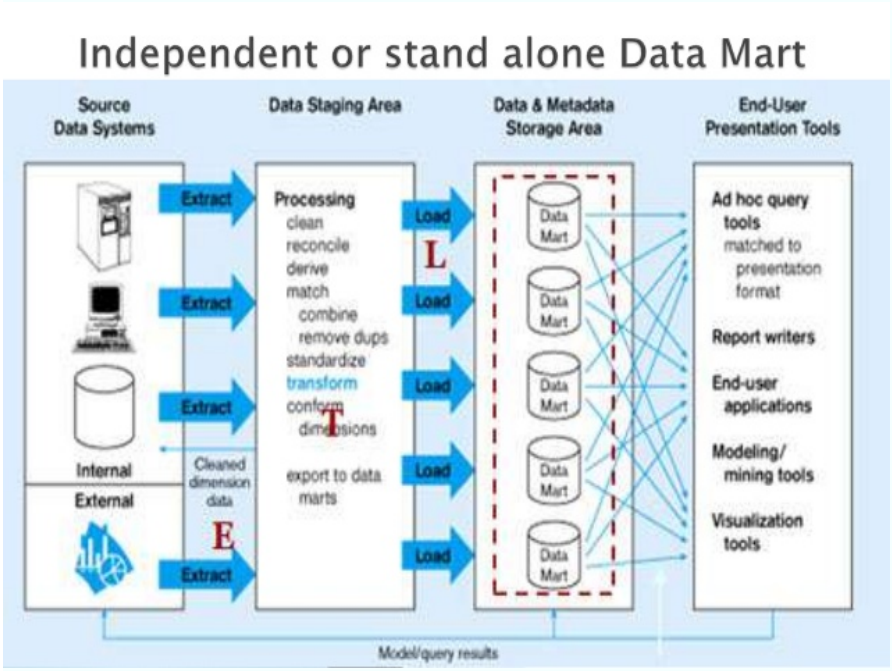
- So lots of hype is generated constantly to keep the topic hot.
- Buzzwords
  - OLAP, Data warehouses, Data Marts, Data Lakes, Business Intelligence,…

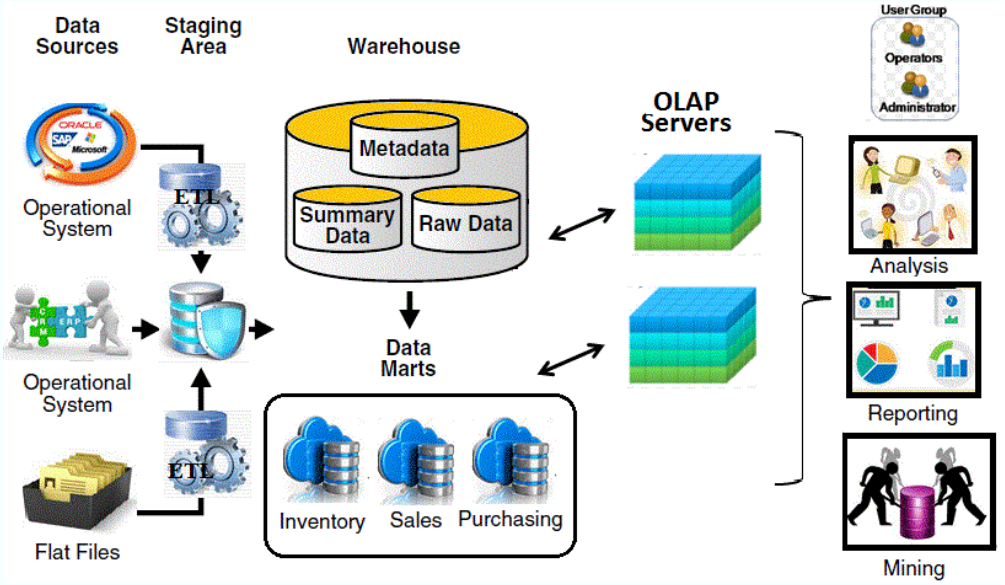


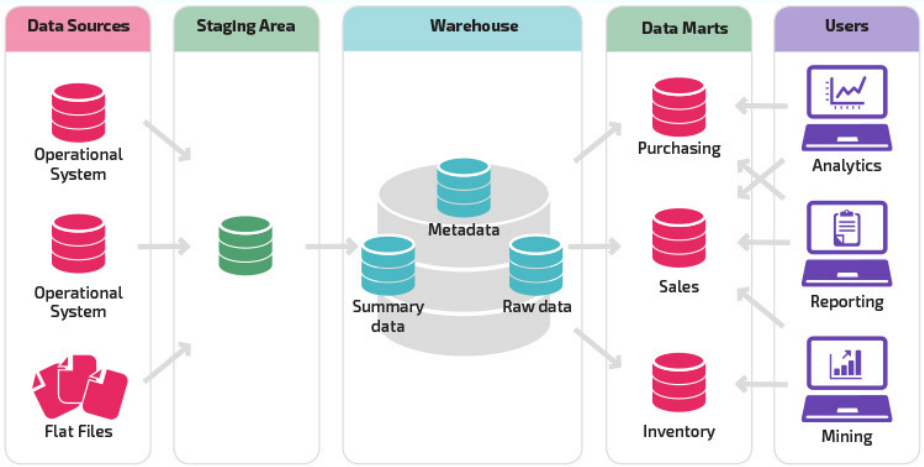




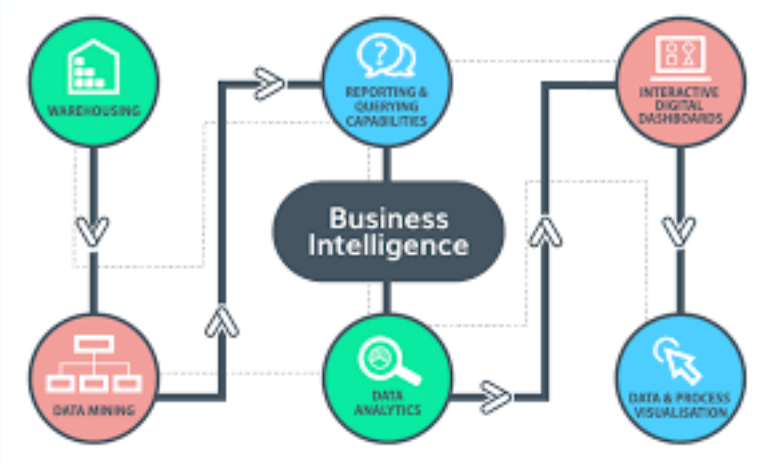
	Most Important Use Group & Use-Cases	Time-to-Market Questions & Solutions	Cost Implementation & Ownership	Users (# & Types)	Data Growth Volume & Variety
Data Lake	Predictive & Advanced Analytics	Weeks - Months	$$$$$		
Data Warehouse	Multi-Purpose Enabler of Operational & Performance Analytics	Hours - Days	$$$$$		
Data Mart	Line of Business Specific Reporting & Analytics	Minutes - Hours	$$$$$		


Independent or stand alone Data Mart
Source Data Systems
Data Staging Area
Data & Metadata Storage Area
End-User Presentation Tools
Extract
Processing
clean
reconcile
derive
match
combine
remove dups
standardize
transform
conform
dimensions
export to data marts
Load
L
T
E
Internal
External
Cleaned dimension data
Data Mart
Ad hoc query tools
matched to presentation format
Report writers
End-user applications
Modeling/ mining tools
Visualization tools
Model/query results

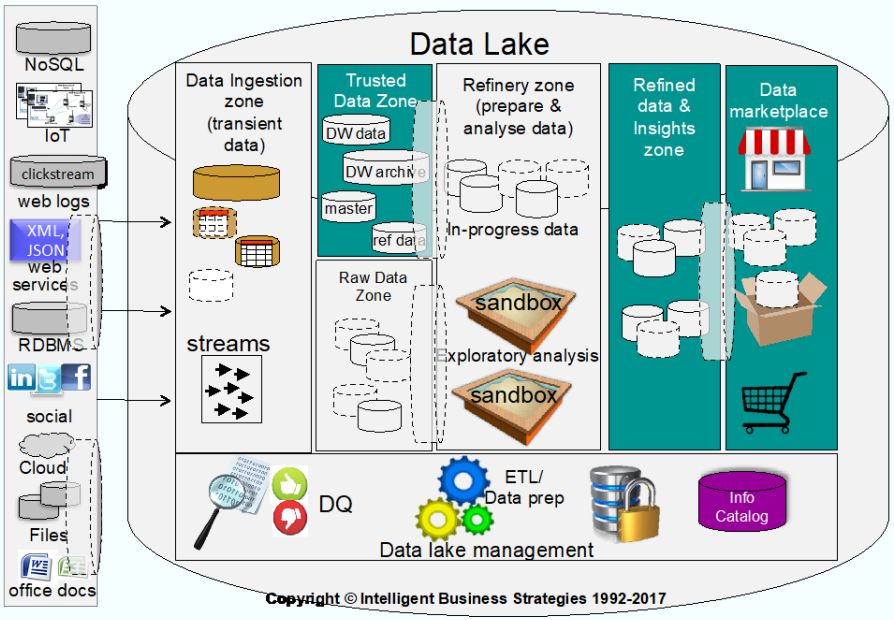

Data Sources
Staging Area
Warehouse
OLAP Servers
User Group
Operators
Administrator
Operational System
Operational System
Flat Files
ETL
ETL
Metadata
Summary Data
Raw Data
Data Marts
Inventory
Sales
Purchasing
Analysis
Reporting
Mining

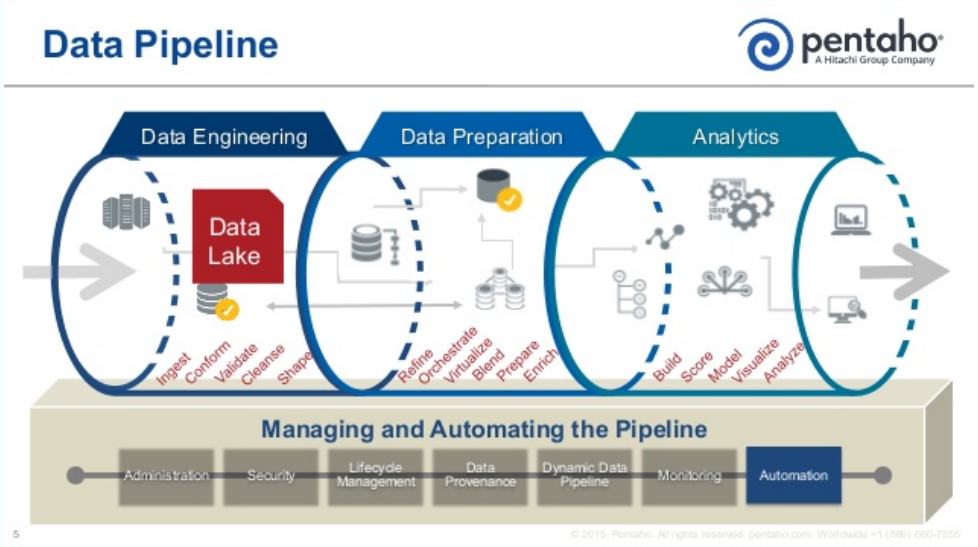

Data Sources
Staging Area
Warehouse
Data Marts
Users
Operational System
Operational System
Flat Files
Metadata
Summary data
Raw data
Purchasing
Sales
Inventory
Analytics
Reporting
Mining

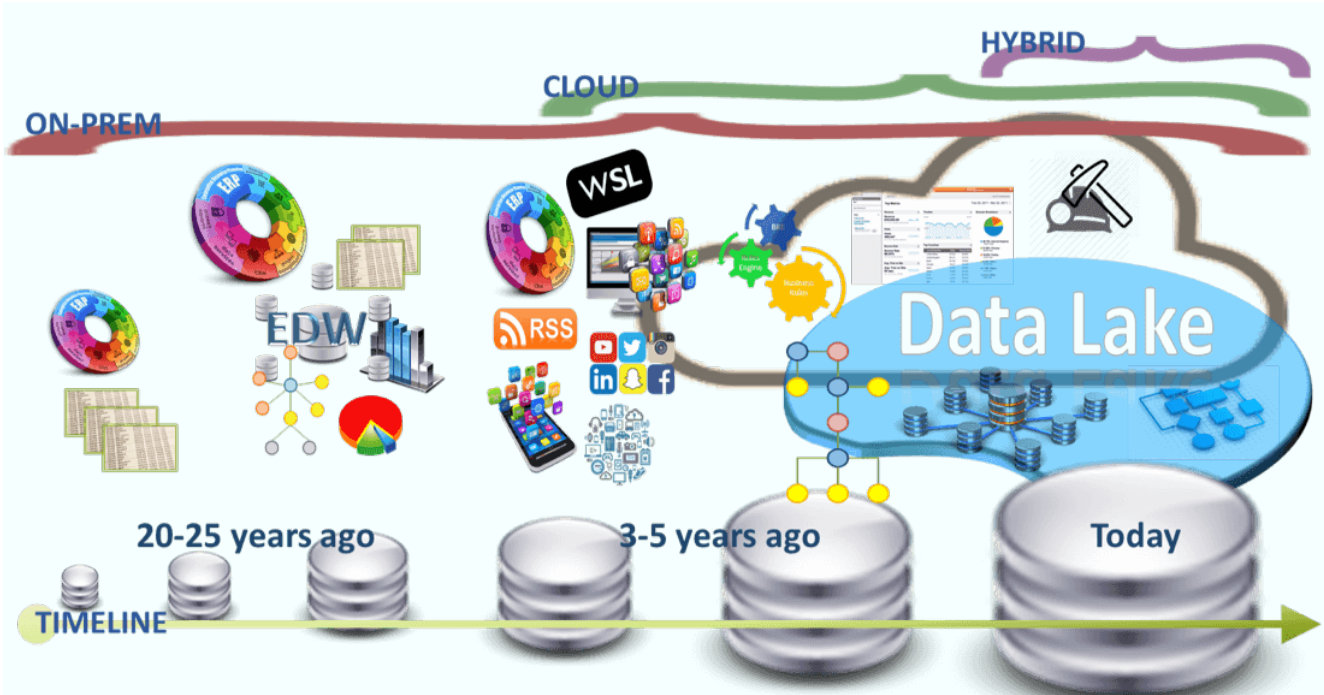

BUSINESS INTELLIGENCE
ERP



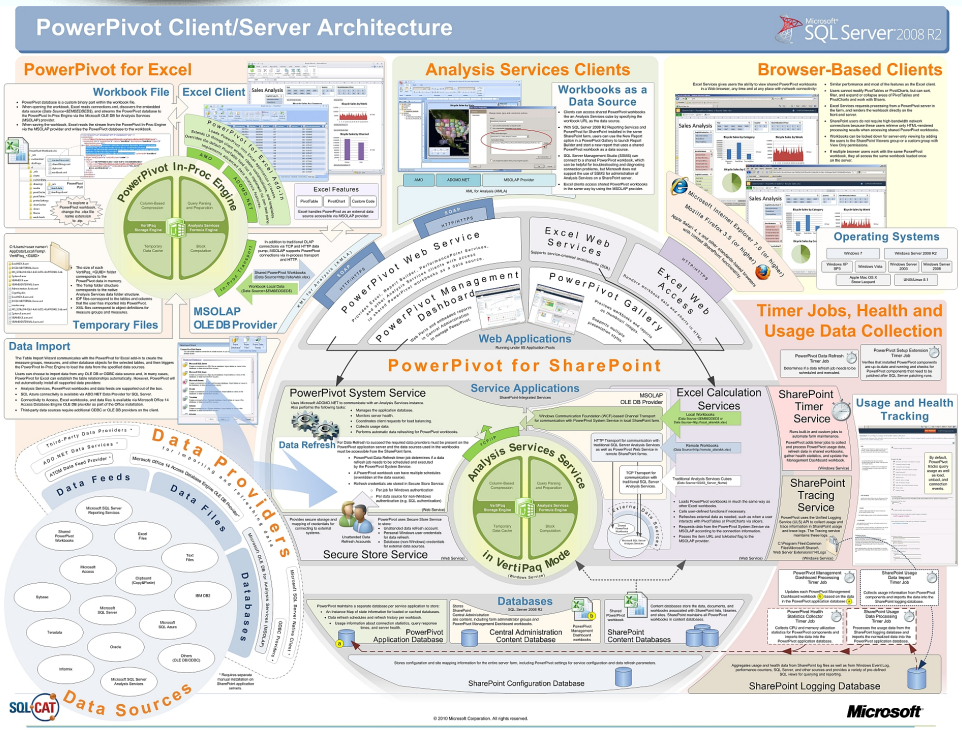
Data Virtualisation Can Help Introduce Agility Into A Data Lake While Reducing Data Copying






HYBRID
CLOUD
ON-PREM
EDW
WSL
RSS
Data Lake
20-25 years ago
3-5 years ago
Today
TIMELINE


Business Intelligence Process
Decision Making
Reporting
Data Analysis
OLAP
Data Mining
Data Warehouse
Extract-Transform-Load
Data Sources


PowerPivot Client/Server Architecture
SQL Server 2008 R2
PowerPivot for Excel
Workbook File
Excel Client
PowerPivot In-Proc Engine
Temporary Files
MSOLAP OLE DB Provider
Data Import
Analysis Services Clients
Workbooks as a Data Source
Browser-Based Clients
Operating Systems
PowerPivot Web Service
PowerPivot Management Dashboard
Excel Web Services
PowerPivot Gallery
Excel Web Access
Web Applications
PowerPivot for SharePoint
Timer Jobs, Health and Usage Data Collection
PowerPivot System Service
Service Applications
Excel Calculation Services
SharePoint Timer Service
Usage and Health Tracking
Data Refresh
Analysis Services Service
in VertiPaq Mode
SharePoint Tracing Service
Secure Store Service
Data Providers
Data Feeds
Data Files
Databases
Data Sources
Databases
PowerPivot Application Database
Central Administration Content Database
SharePoint Content Databases
SharePoint Configuration Database
SharePoint Logging Database
SQL CAT
Microsoft

Introduction

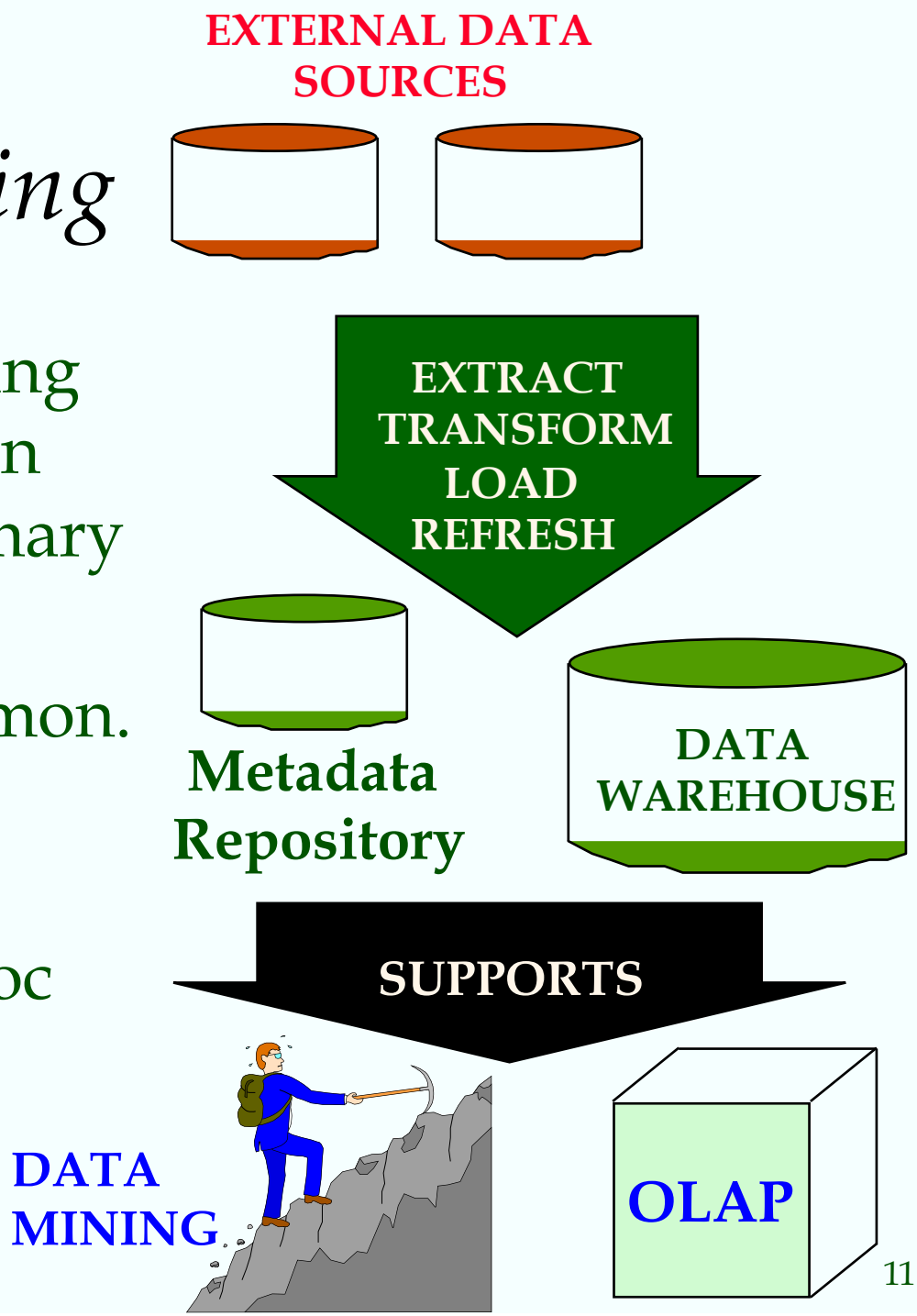
- Increasingly, organizations are analyzing current and historical data to identify useful patterns and support business strategies.
- Emphasis is on complex, interactive, exploratory analysis of very large datasets created by integrating data from across all parts of an enterprise; data is fairly static.
  - Contrast such **On-Line Analytic Processing (OLAP)** with traditional **On-line Transaction Processing (OLTP):** mostly long queries, instead of short update Xacts.

Three Complementary Trends

- Data Warehousing: Consolidate data from many sources in one large repository.
  - Loading, periodic synchronization of replicas.
  - Semantic integration.
- OLAP:
  - Complex SQL queries and views.
  - Queries based on spreadsheet-style operations and "multidimensional" view of data.
  - Interactive and "online" queries.
- Data Mining: Exploratory search for interesting trends and anomalies.

Data Warehousing

- Integrated data spanning long time periods, often augmented with summary information.
- Several terabytes common.
- Interactive response times expected for complex queries; ad-hoc updates uncommon.



Warehousing Issues

- Semantic Integration: When getting data from multiple sources, must eliminate mismatches, e.g., different currencies, schemas.
- Heterogeneous Sources: Must access data from a variety of source formats and repositories.
  - Replication capabilities can be exploited here.
- Load, Refresh, Purge: Must load data, periodically refresh it, and purge too-old data.
- Metadata Management: Must keep track of source, loading time, and other information for all data in the warehouse.

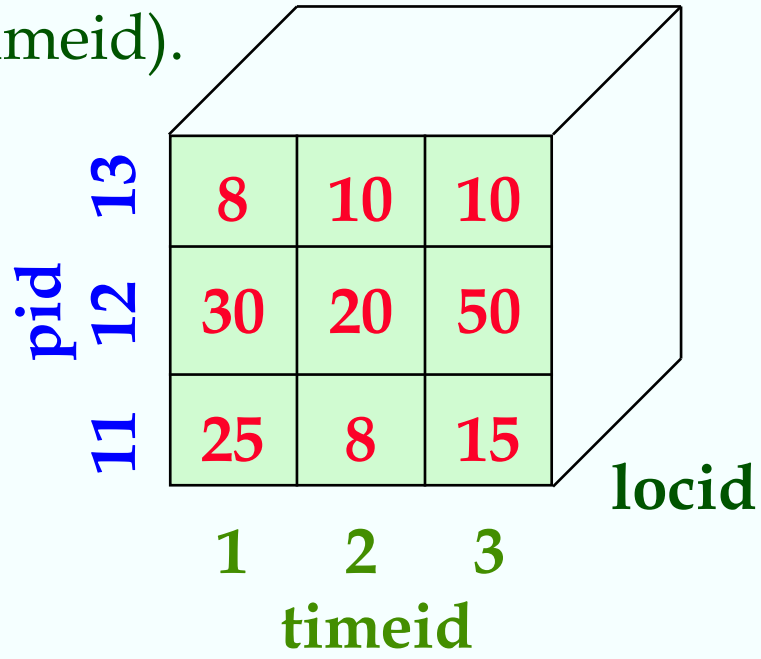
Data Quality; Data Cleaning

- Needs metrics; needs to be addressed proactively
  - Redesign processes for data gathering.
  - Feedback loops
  - Assign $ cost to low quality data.
  - Assign people responsible for data quality
- Key problems:
  - Erraneous and missing data
  - Truncation and censoring
  - Entity resolution/deduplication/record linkage

Multidimensional Data Model

- Collection of numeric measures, which depend on a set of dimensions.
  - E.g., measure **Sales**, dimensions **Product** (key: pid), **Location** (locid), and **Time** (timeid).

Slice locid=1 is shown:



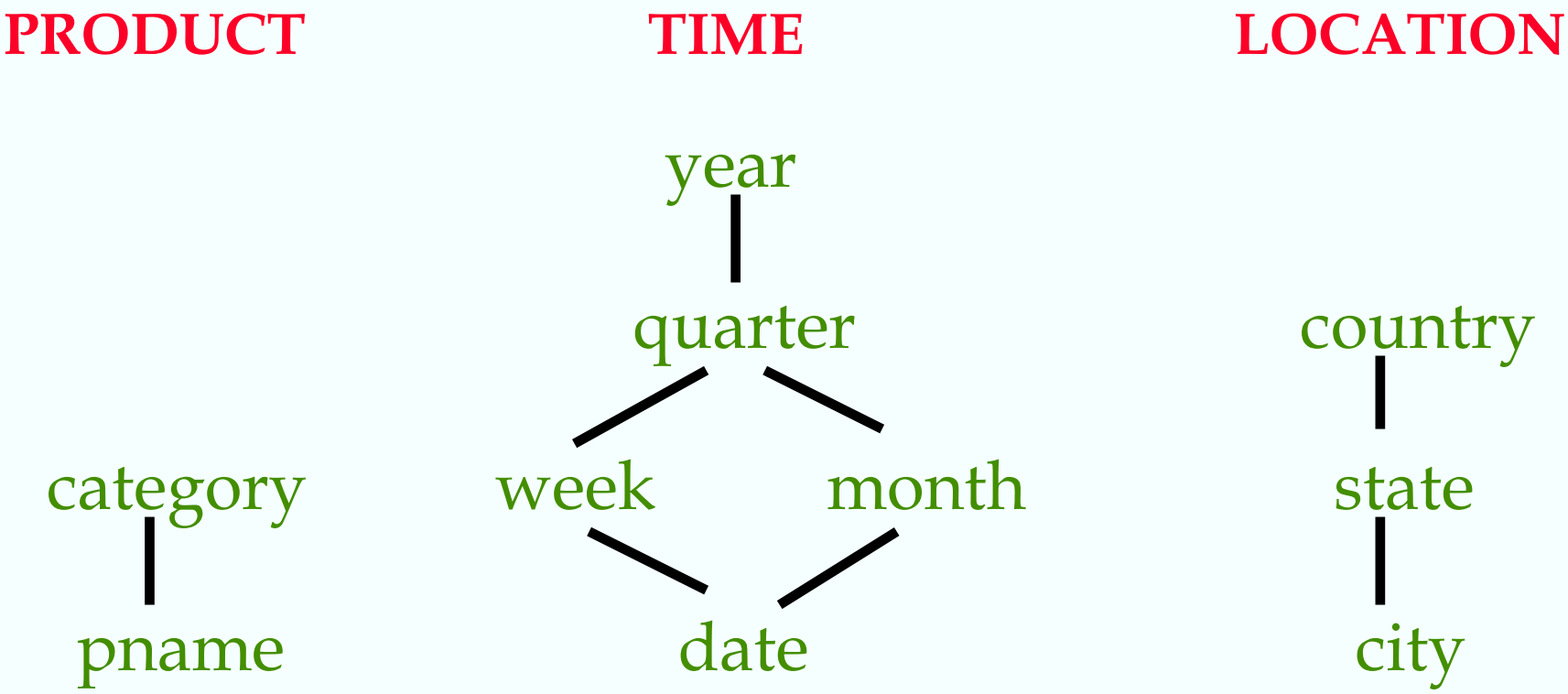
pid	timeid	locid	sales
11	1	1	25
11	2	1	8
11	3	1	15
12	1	1	30
12	2	1	20
12	3	1	50
13	1	1	8
13	2	1	10
13	3	1	10
11	1	2	35

MOLAP vs ROLAP

- Multidimensional data can be stored physically in a (disk-resident, persistent) array; called MOLAP systems. Alternatively, can store as a relation; called ROLAP systems.
- The main relation, which relates dimensions to a measure, is called the fact table. Each dimension can have additional attributes and an associated dimension table.
  - E.g., **Products(pid, pname, category, price)**
  - Fact tables are *much* larger than dimensional tables.

Dimension Hierarchies

- For each dimension, the set of values can be organized in a hierarchy:



OLAP Queries

- Influenced by SQL and by spreadsheets.
- A common operation is to aggregate a measure over one or more dimensions.
  - Find total sales.
  - Find total sales for each city, or for each state.
  - Find top five products ranked by total sales.
- Roll-up: Aggregating at different levels of a dimension hierarchy.
  - E.g., Given total sales by city, we can roll-up to get sales by state.

OLAP Queries

- Drill-down: The inverse of roll-up.
  - E.g., Given total sales by state, can drill-down to get total sales by city.
  - E.g., Can also drill-down on different dimension to get total sales by product for each state.
- Pivoting: Aggregation on selected dimensions.
  - E.g., Pivoting on Location and Time yields this **cross-tabulation**:

	WI	CA	Total
1995	63	81	144
1996	38	107	145
1997	75	35	110
Total	176	223	339

- Slicing and Dicing: Equality and range selections on one or more dimensions.

Comparison with SQL Queries

- The cross-tabulation obtained by pivoting can also be computed using a collection of SQLqueries:

```
SELECT SUM(S.sales)
FROM   Sales S, Times T, Locations L
WHERE  S.timeid=T.timeid AND S.timeid=L.timeid
GROUP BY T.year, L.state
```

```
SELECT SUM(S.sales)
FROM   Sales S, Times T
WHERE  S.timeid=T.timeid
GROUP BY T.year
```

```
SELECT SUM(S.sales)
FROM   Sales S, Location L
WHERE  S.timeid=L.timeid
GROUP BY L.state
```

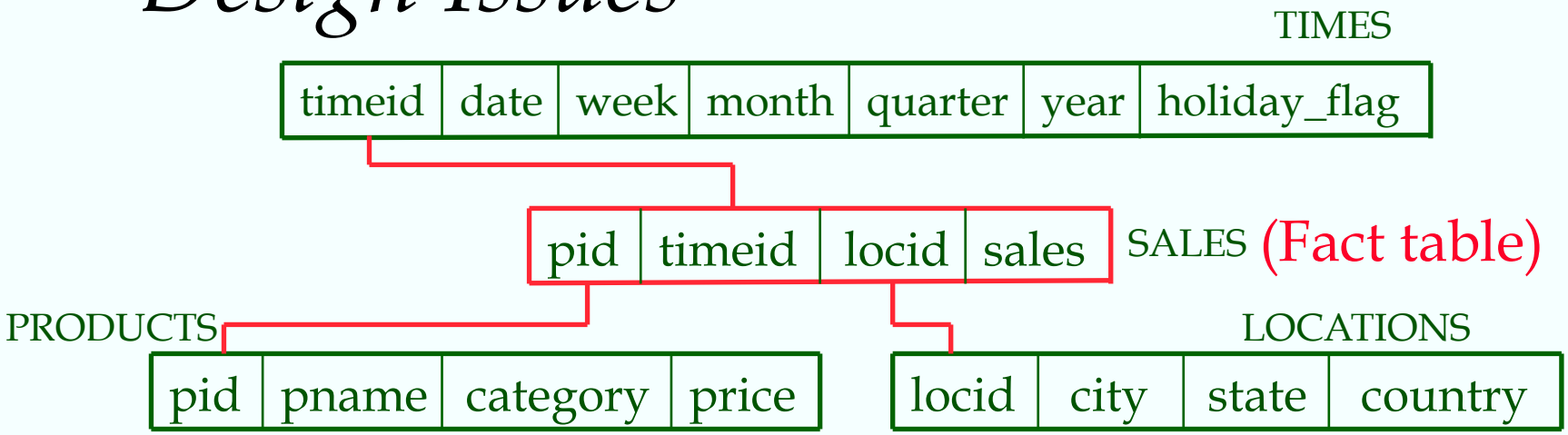
The CUBE Operator

- Generalizing the previous example, if there are k dimensions, we have 2^k possible SQL GROUP BY queries that can be generated through pivoting on a subset of dimensions.
- CUBE pid, locid, timeid BY SUM Sales
  - Equivalent to rolling up Sales on all eight subsets of the set {pid, locid, timeid}; each roll-up corresponds to an SQL query of the form:

```
SELECT SUM(S.sales)
FROM   Sales S
GROUP BY grouping-list
```

Lots of work on optimizing the CUBE operator!

Design Issues

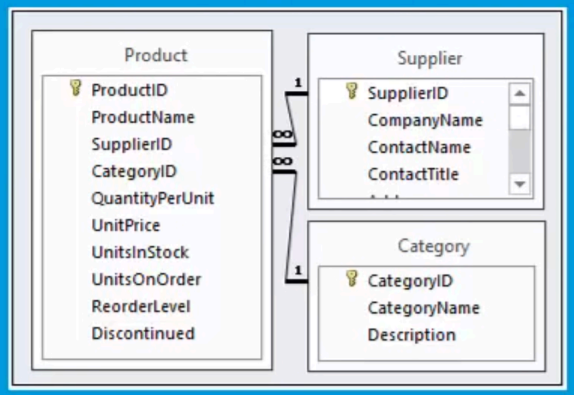


- Fact table in BCNF; dimension tables un-normalized.
  - Dimension tables are small; updates/inserts/deletes are rare. So, anomalies less important than query performance.
- This kind of schema is very common in OLAP applications, and is called a star schema; computing the join of all these relations is called a star join.

Excel Tutorial: What is Business Intelligence and an OLAP Cube? | ExcelCentral.com

MDX (multi-dimensional expressions)

Relational Database



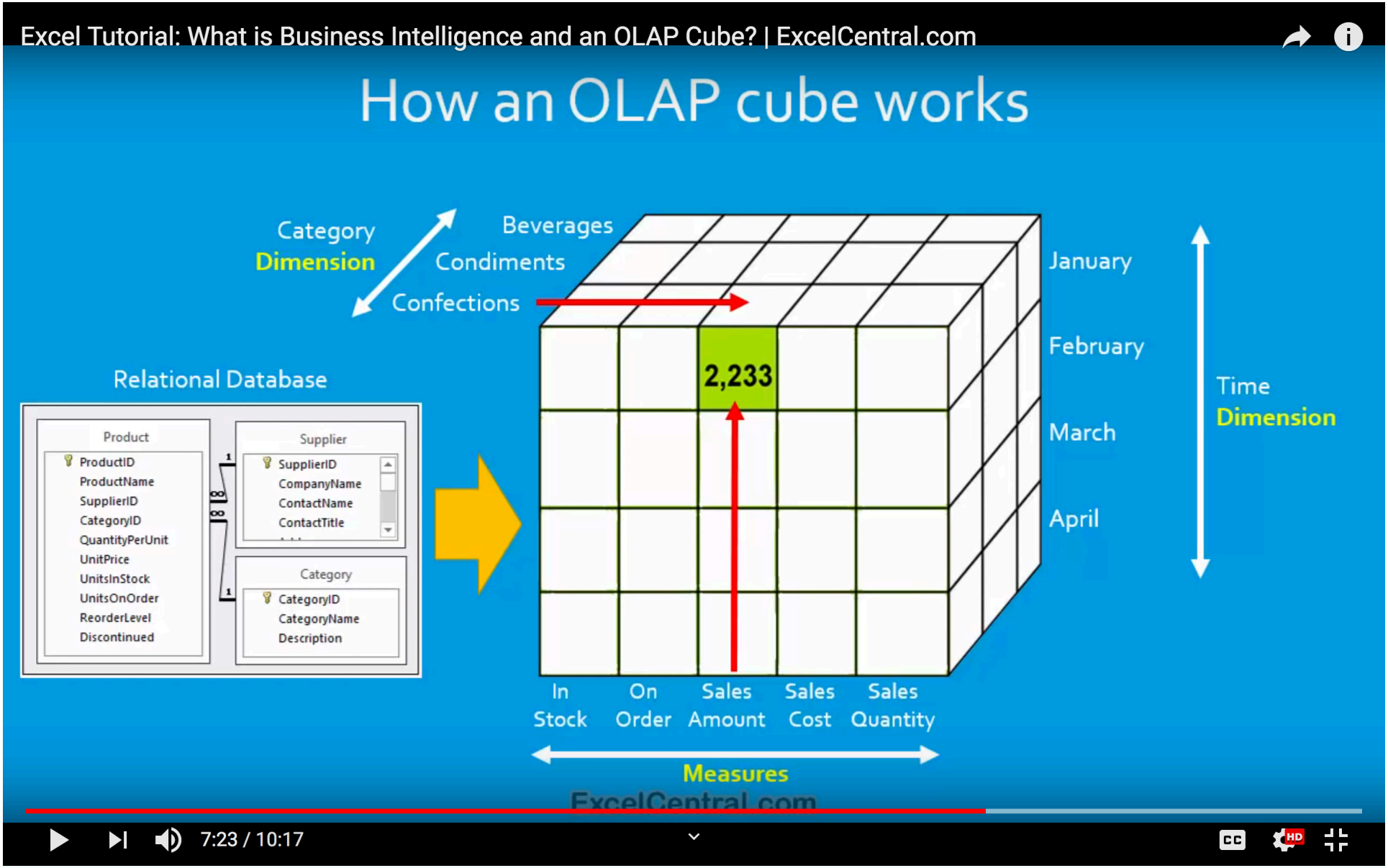
OLAP Cube

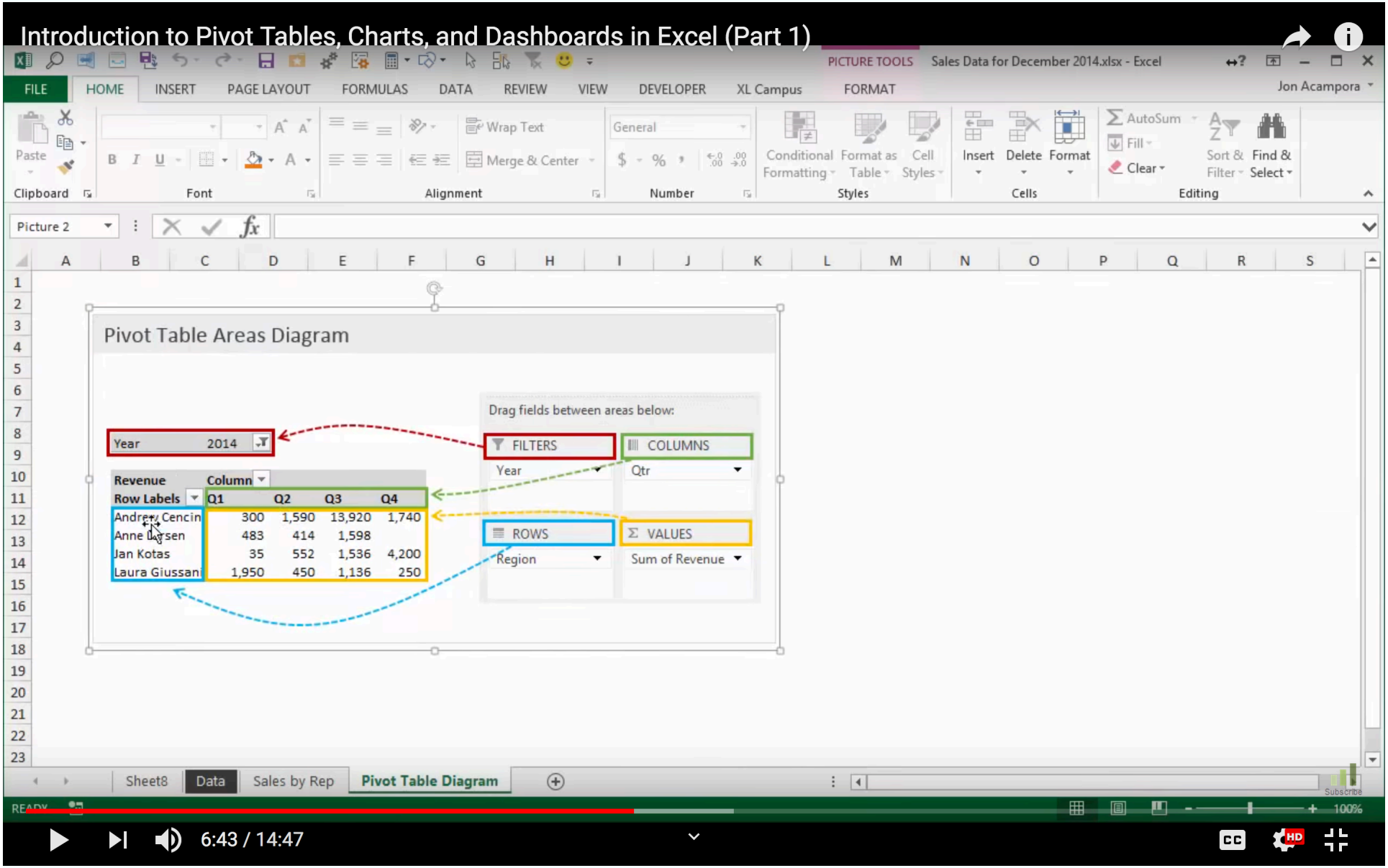
OLAP Pivot Table

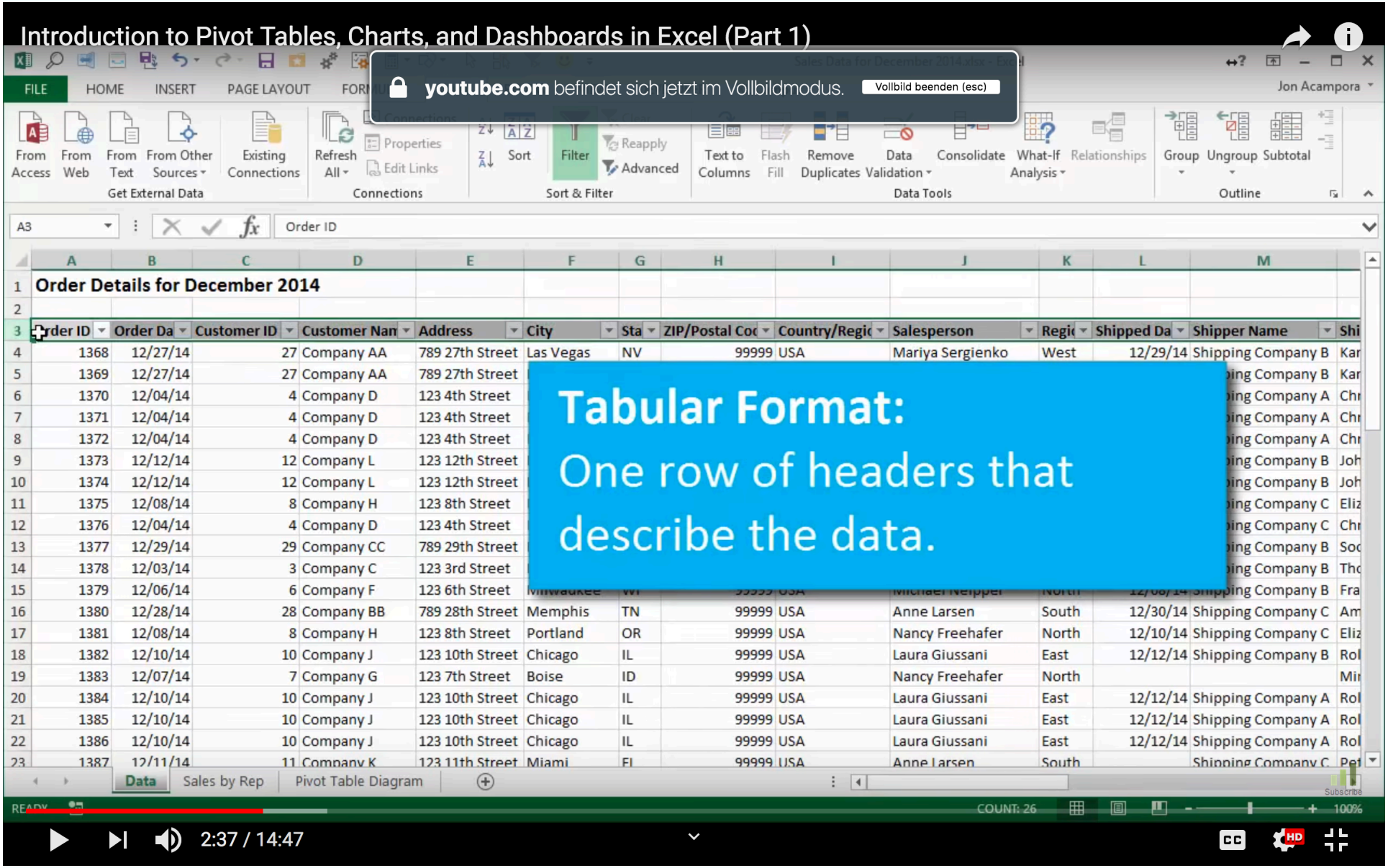
	A	B	C
1			
2	Row Labels	Sum of UnitsInStock	Sum of UnitsOnOrder
3	Beverages	559	60
4	Condiments	507	170
5	Confections	386	180
6	Dairy Products	393	140
7	Grains/Cereals	308	90
8	Meat/Poultry	165	0
9	Produce	100	20
10	Seafood	701	120
11	Grand Total	3119	780

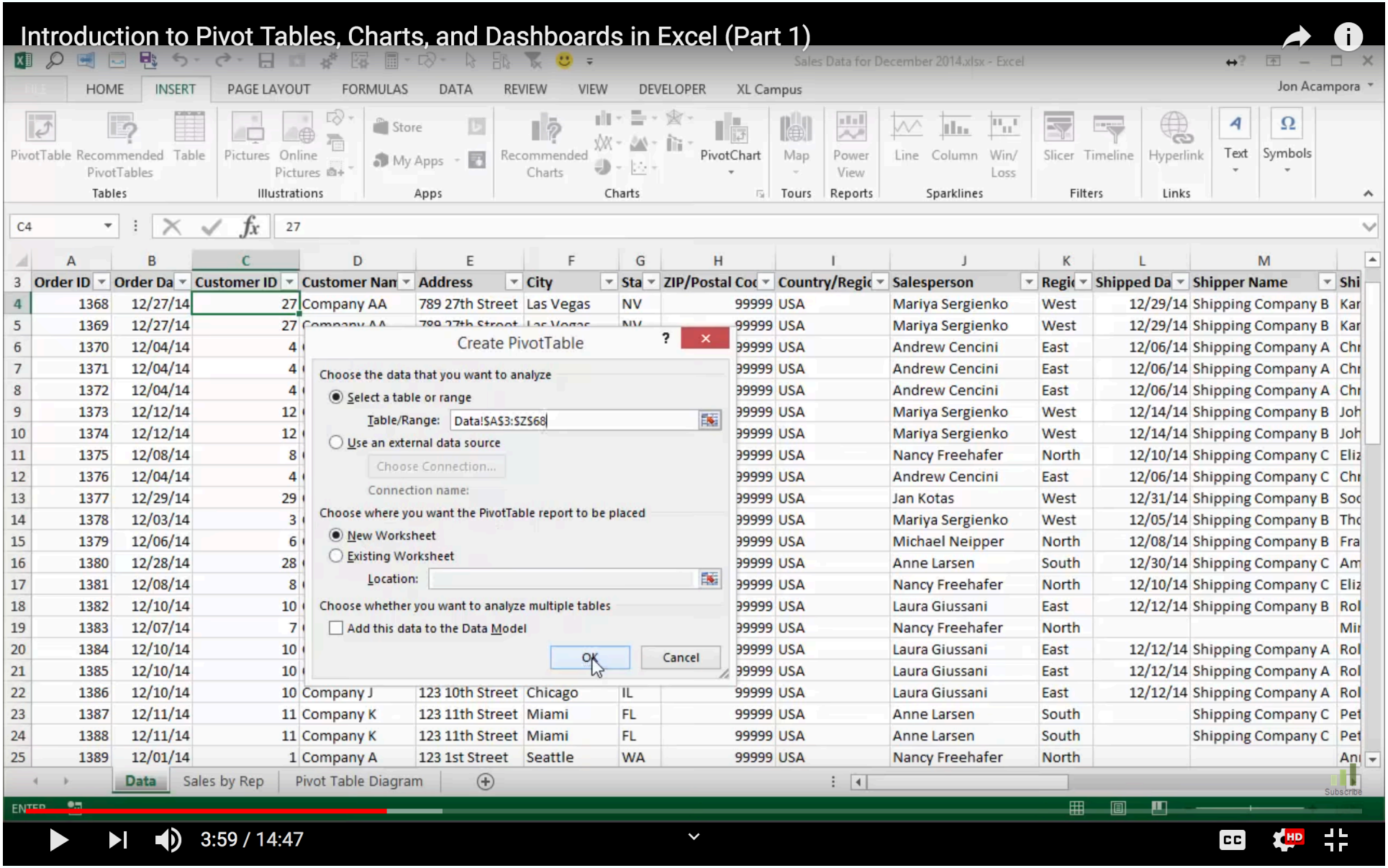
ExcelCentral.com

9:00 / 10:17


Excel Tutorial: What is Business Intelligence and an OLAP Cube? | ExcelCentral.com
How an OLAP cube works
Category
Dimension
Beverages
Condiments
Confections
Relational Database
Product
ProductID
ProductName
SupplierID
CategoryID
QuantityPerUnit
UnitPrice
UnitsInStock
UnitsOnOrder
ReorderLevel
Discontinued
Supplier
SupplierID
CompanyName
ContactName
ContactTitle
Category
CategoryID
CategoryName
Description
2,233
January
February
March
April
Time
Dimension
In Stock
On Order
Sales Amount
Sales Cost
Sales Quantity
Measures
ExcelCentral.com
7:23 / 10:17


Introduction to Pivot Tables, Charts, and Dashboards in Excel (Part 1)
Sales Data for December 2014.xlsx - Excel
Pivot Table Areas Diagram
Drag fields between areas below:
FILTERS
COLUMNS
ROWS
VALUES
Year
Qtr
Region
Sum of Revenue
Year 2014
Revenue
Column
Row Labels
Q1
Q2
Q3
Q4
Andrew Cencin 300 1,590 13,920 1,740
Anne Larsen 483 414 1,598
Jan Kotas 35 552 1,536 4,200
Laura Giussani 1,950 450 1,136 250
Sheet8
Data
Sales by Rep
Pivot Table Diagram
6:43 / 14:47

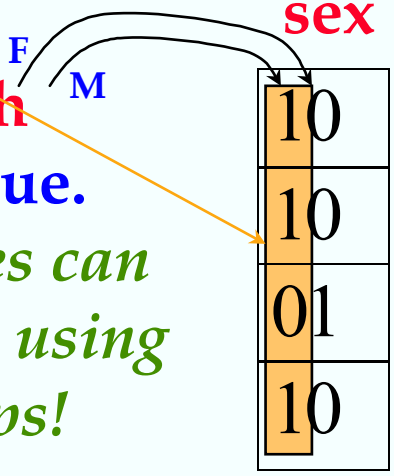

Introduction to Pivot Tables, Charts, and Dashboards in Excel (Part 1)
youtube.com befindet sich jetzt im Vollbildmodus.
Vollbild beenden (esc)
Tabular Format:
One row of headers that describe the data.
2:37 / 14:47


Introduction to Pivot Tables, Charts, and Dashboards in Excel (Part 1)
Sales Data for December 2014.xlsx - Excel
HOME
INSERT
PAGE LAYOUT
FORMULAS
DATA
REVIEW
VIEW
DEVELOPER
XL Campus
Jon Acampora
Create PivotTable
Choose the data that you want to analyze
Select a table or range
Table/Range: Data!$A$3:$Z$68
Use an external data source
Choose Connection...
Connection name:
Choose where you want the PivotTable report to be placed
New Worksheet
Existing Worksheet
Location:
Choose whether you want to analyze multiple tables
Add this data to the Data Model
OK
Cancel
Data
Sales by Rep
Pivot Table Diagram
3:59 / 14:47

Implementation Issues

- New indexing techniques: Bitmap indexes, Join indexes, array representations, compression, precomputation of aggregations, etc.
- E.g., Bitmap index:

Bit-vector:
1 bit for each possible value.
Many queries can be answered using bit-vector ops!



sex

10
10
01
10

custid name sex rating

112	Joe	M	3
115	Ram	M	5
119	Sue	F	5
112	Woo	M	4

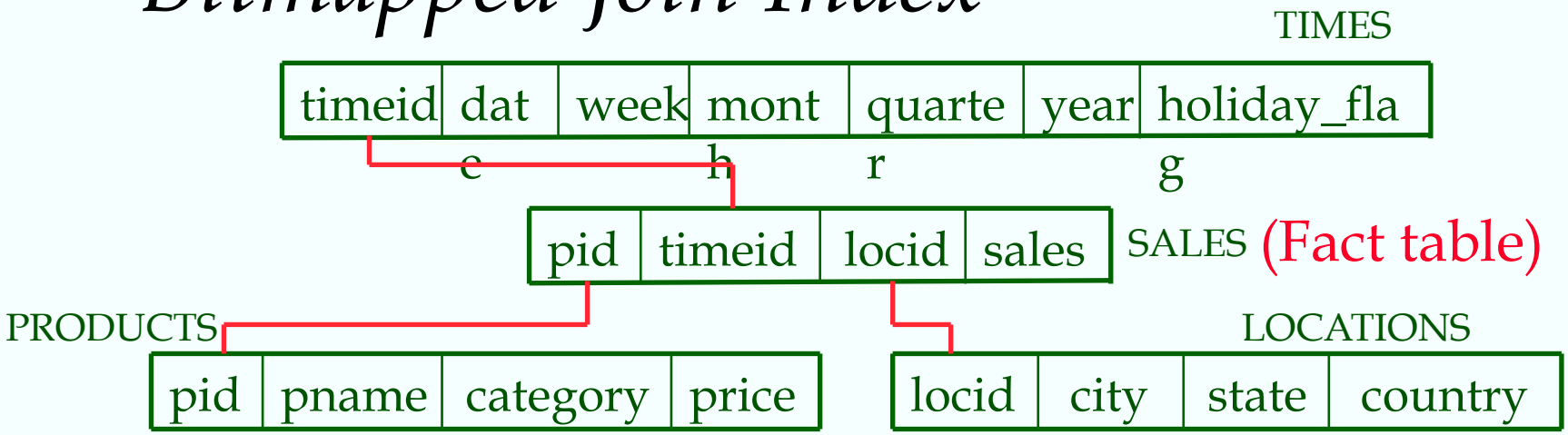
rating

00100
00001
00001
00010

Join Indexes

- Consider the join of Sales, Products, Times, and Locations, possibly with additional selection conditions (e.g., country="USA").
  - A join index can be constructed to speed up such joins. The index contains [s,p,t,l] if there are tuples (with sid) s in Sales, p in Products, t in Times and l in Locations that satisfy the join (and selection) conditions.
- Problem: Number of join indexes can grow rapidly.
  - A variation addresses this problem: For each column with an additional selection (e.g., country), build an index with [c,s] in it if a dimension table tuple with value c in the selection column joins with a Sales tuple with sid s; if indexes are bitmaps, called bitmapped join index.

Bitmapped Join Index

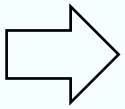


- Consider a query with conditions price=10 and country="USA". Suppose tuple (with sid) s in Sales joins with a tuple p with price=10 and a tuple l with country ="USA". There are two join indexes; one containing [10,s] and the other [USA,s].
- Intersecting these indexes tells us which tuples in Sales are in the join and satisfy the given selection.

Column Stores

- ROLAP approach in which relations are stored column by column, rather than row by row.

sid	sname
1	Alice
2	Bob
3	Charles



sid
1
2
3

sname
Alice
Bob
Charles

- Order and duplicates matter! Like (lossless) vertical decomposition with implicit rowid as key.
- Optimizations: column compression usually works better than table compression.
- Simple idea, but very fashionable right now. (HP Vertica, SAP HANA, ⋯)

Row Stores vs Column Stores

- RDBMS are traditionally implemented as row stores. Can be implemented as column stores.
- Optimize for OLTP => row store
  - Optimizes date locality for updating. A tuple is in one place only
- Optimize for analytics => column store
  - Assuming each query uses only some of the columns => less data to scan