Data Integration: From The Beginning to Today

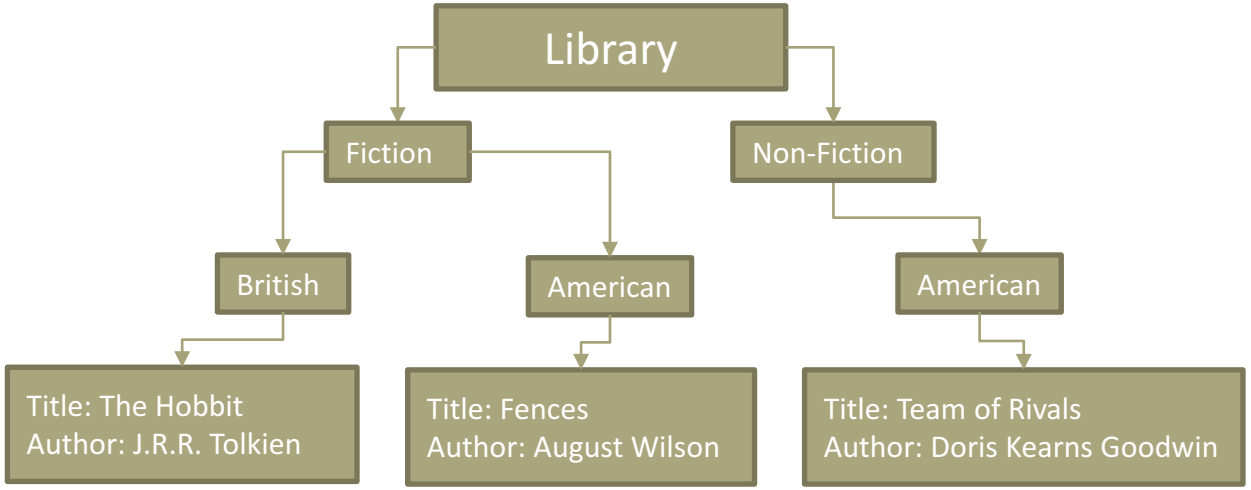
Mary Roth
IBM Research – Almaden
January 2018



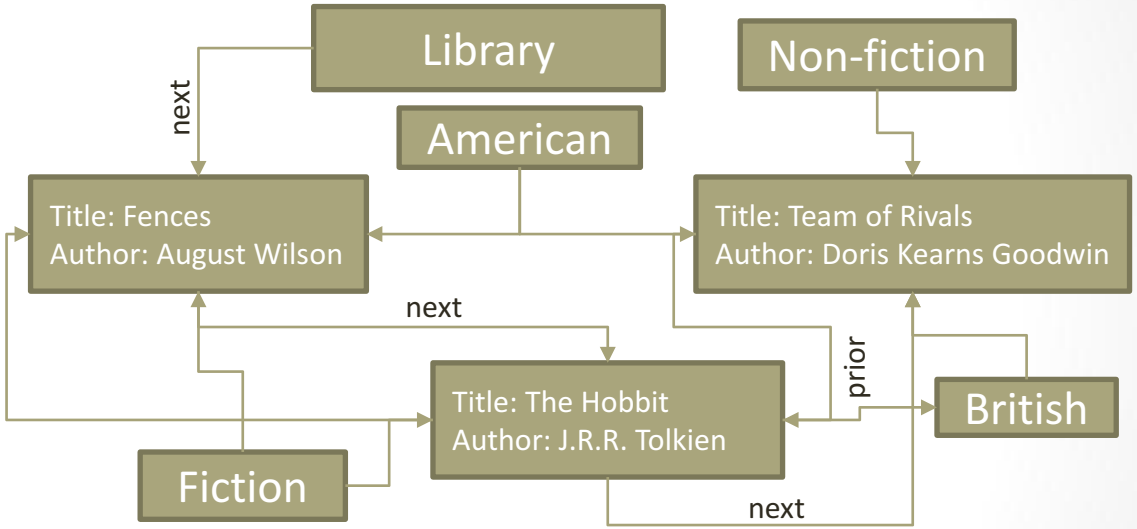
In the beginning, Codd made the relational database…

- In the world before Codd (~1960s), these data models prevailed:

Hierarchical (e.g. IMS)



Network Model (e.g. CODASYL)



- But these models did not provide *data independence*;
  - If the data layout changed, the application had to change
  - If you wanted to change the layout, you often had to bring the whole system down
  - Changes had to occur over scheduled system down time.
    - Slow! Annoying! Expensive!

The relational database ***revolutionized*** how we electronically store and retrieve data

And Codd saw that the relational database was good.

- It provided a structured, organized and application independent way to store data. [Codd 1970]

Books in a Library…

Title	Author	Dewey	Category
Fences	Wilson, August	812.54	Drama
The Hobbit	Tolkien, J. R. R.	823.912	Fiction
Team of Rivals	Goodwin, Doris Kearns	973.7092 B 22	Non-fiction

Or books in a Bookstore…

Title	Author	ISBN	Price
Fences	Wilson, August	0452264014	8.47
The Hobbit	Tolkien, J. R. R.	9780007440849	10.55
Team of Rivals	Goodwin, Doris Kearns	0743270754	13.42

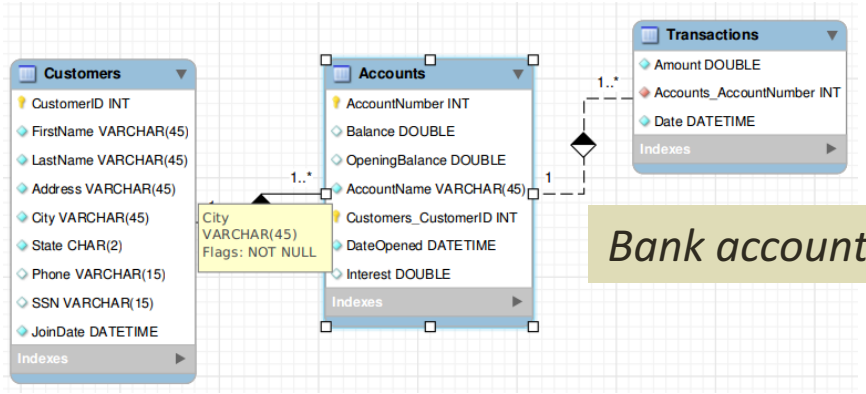
Or even groceries in a Grocery Store!

Name	Aisle	UPC	Price
Chobani Yogurt	Dairy	894700010014	1.00
Jiffy Peanut Butter	12	051500255162	5.49
Welch's Fruit Snacks	14	034856050926	4.29

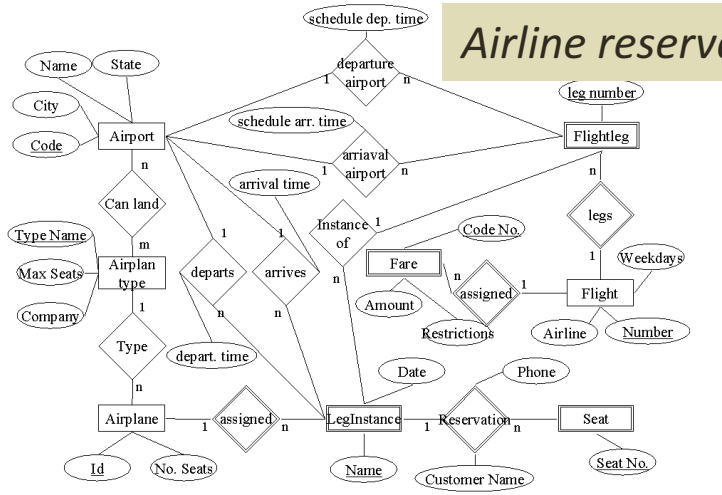
Key Advantages:

- Data could be restructured without affecting applications
- Data could be quickly and efficiently retrieved via arbitrarily complex queries

Over time, commercial enterprises began to rely on relational databases to run all aspects of their business.

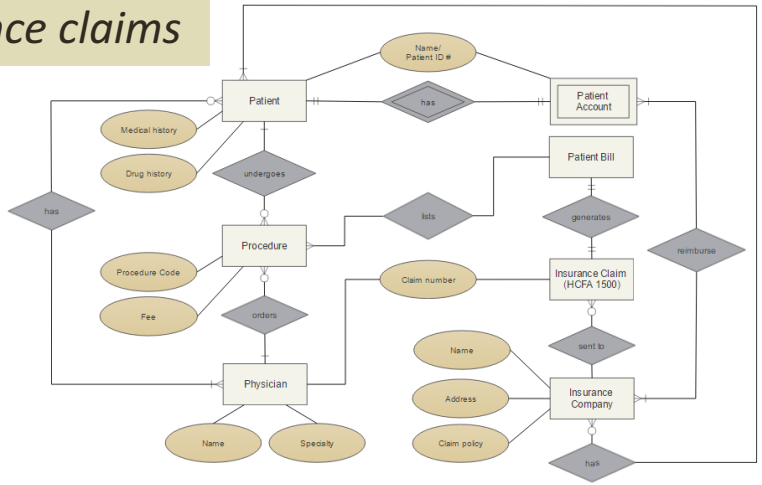


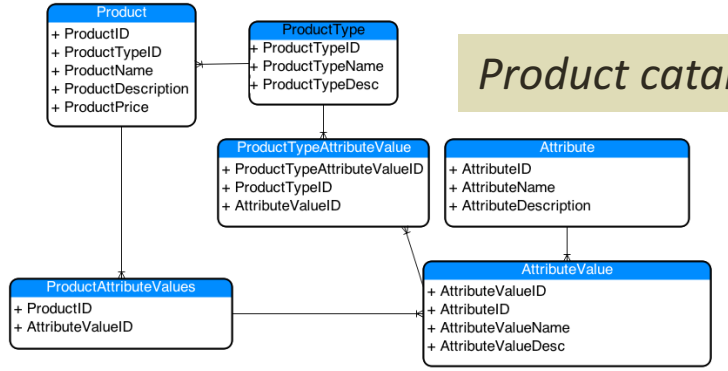
Bank accounts



Airline reservations

Insurance claims





Product catalogs, and more…

Eventually, this led to new desires to combine data from multiple different places, e.g.,:

- *How can I run analysis across all of the bank accounts at all of my branches?*
- *How can I find out if a book is at any library or book store near me?*

These desires led to new challenges in *data integration.*

UC-Santa Barbara Branch

Title	Author	Dewey	Category
Fences	Wilson, August	812.54	Drama
The Hobbit	Tolkien, J. R. R.	823.912	Fiction
Team of Rivals	Goodwin, Doris Kearns	973.7092 B 22	Non-fiction

UCLA Bookstore

Title	Author	ISBN	Price
Fences	Wilson, August	0452264014	8.47
The Hobbit	Tolkien, J. R. R.	9780007440849	10.55
Team of Rivals	Goodwin, Doris Kearns	0743270754	13.42



1. How do I query and combine data that is physically stored in different places?

*2. How do I query and combine data that is **described** differently in different places?*

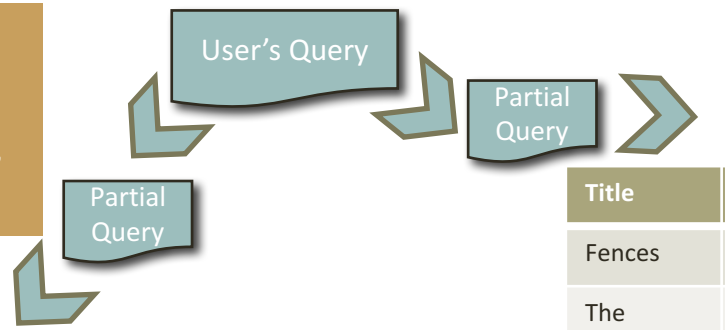
3. How do I query and combine data that is represented differently in different places?

UCLA Branch

Title	Last name	First name	Dewey	Category
Fences	Wilson	August	812.54	Drama
The Hobbit	Tolkien	J.R.R.	823.912	Fiction
Team of Rivals	Kearnes Goodwin	Doris	973.7092 B 22	Non-fiction

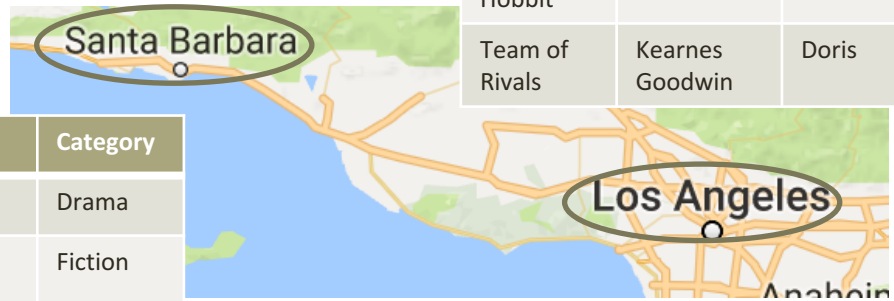
Challenge 1: How do I query data physically stored in different places?

Approach 1: Data Federation
leave the data in place and move the query to the data. [Roth, et. al, 1996, Tomasic, et. al, 1996]



UCLA Branch

Title	Last name	First name	Dewey	Category
Fences	Wilson	August	812.54	Drama
The Hobbit	Tolkien	J.R.R.	823.912	Fiction
Team of Rivals	Kearnes Goodwin	Doris	973.7092 B 22	Non-fiction



UC-Santa Barbara Branch

Title	Author	Dewey	Category
Fences	Wilson, August	812.54	Drama
The Hobbit	Tolkien, J. R. R.	823.912	Fiction
Team of Rivals	Goodwin, Doris Kearns	973.7092 B 22	Non-fiction

UC Library Catalog

Title	Author	Dewey	Category
Fences	Wilson, August	812.54	Drama
The Hobbit	Tolkien, J. R. R.	823.912	Fiction
Team of Rivals	Goodwin, Doris Kearns	973.7092 B 22	Non-fiction

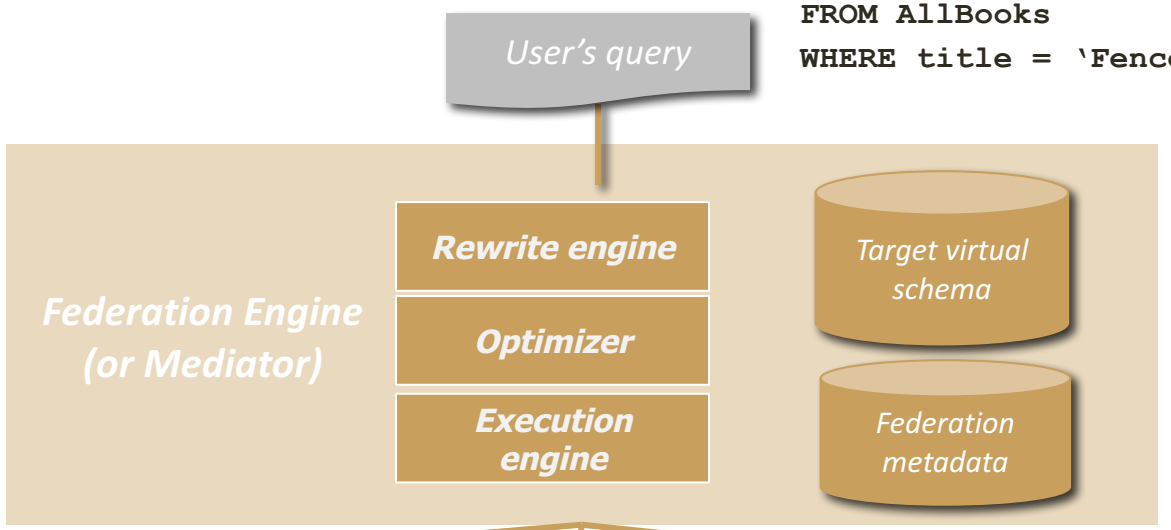
Dewey	Location	#
812.54	UCLA	1
812.54	UCSB	2
823.912	UCLA	2
823.912	UCSB	1
973.7092 B 22	UCLA	3
973.7092 B 22	UCSB	2

Approach 2: Data Warehousing
Move the data to a single place and query it there. [Chaudhuri, et. al, 1997]

Data Federation: Move the query to the data.

Mediation, Virtual Integration, Data Integration, Lazy Integration

```
SELECT title, author, ISBN
FROM AllBooks
WHERE title = 'Fences'
```




```
SELECT title, first+last
FROM UCLABooks
WHERE title = 'Fences'
```

DB2 Wrapper

MySQL Wrapper

REST wrapper

GET http://bookstore.com/api/books/book?title='Fences'

UCLA Database

UCSB Database

```
SELECT title, author
FROM UCSBBooks
WHERE title = 'Fences'
```

UCLA Bookstore

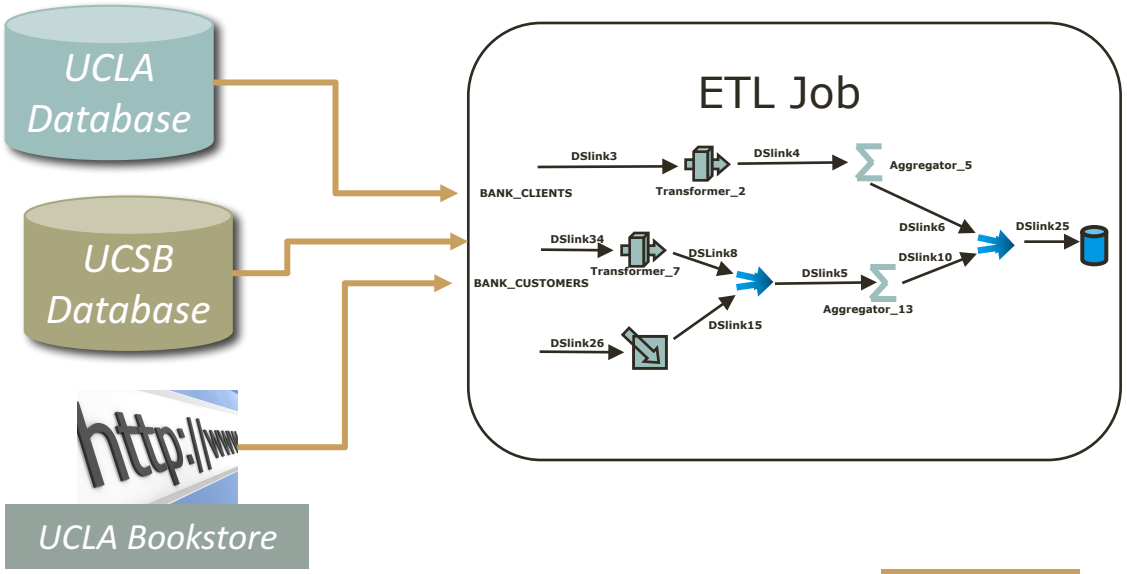
Title	Last name	First name	Dewey	Category
Fences	Wilson	August	812.54	Drama
…		.		

Title	Author	Dewey	Category
Fences	Wilson, August	812.54	Drama
…			

Title	Author	ISBN	Price
Fences	Wilson, August	0452264014	8.47
The Hobbit	Tolkien, J. R. R.	9780007440849	10.55
Team of Rivals	Goodwin, Doris Kearns	0743270754	13.42

Data Warehousing: consolidate the data in one central location.

Materialization, Data Exchange, ETL (Extract, Transform, Load), Eager Integration



User's query

```
SELECT title, author, ISBN,
location
FROM AllBooks A, Locations L
WHERE title = 'Fences'
AND A.ISBN=L.ISBN
```

Target Warehouse Schema

Warehouse Data

AllBooks

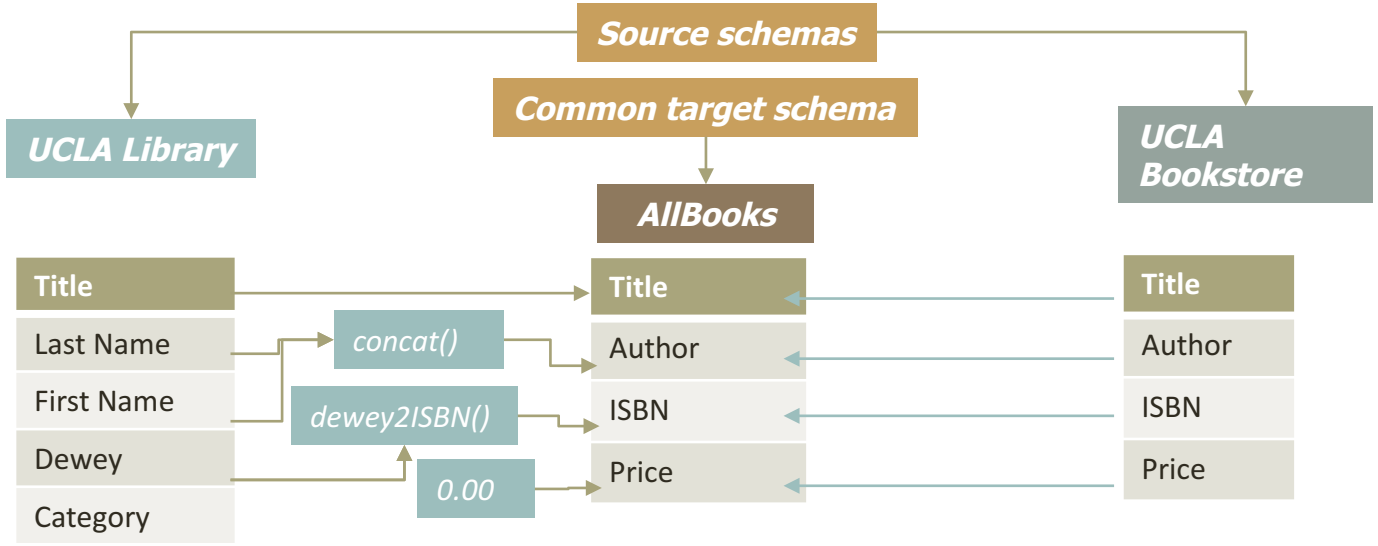
Title	Author	ISBN	Price
Fences	Wilson, August	0452264014	8.47
The Hobbit	Tolkien, J. R. R.	9780007440849	10.55
Team of Rivals	Goodwin, Doris Kearns	0743270754	13.42

Locations

ISBN	Location	#
0452264014	UCLA	1
0452264014	UCSB	2
0452264014	UCLABookStore	2

Challenge 2: How do I query data that is *described* differently in different sources?

Declarative schema mappings describe the relationships between the schemas of heterogeneous data sources [Miller, et. al, 2000]




```
SELECT TITLE, CONCAT(FIRSTNAME, LASTNAME)
       DEWEY2ISBN(DEWEY), 0.00
FROM UCLALIBRARY
```

GET http://bookstore.com/api/books/book

Schema mappings can be compiled into code to restructure the source data to a unified target schema

Challenge 3: How do I query data that is *represented* differently in different sources?

Data can be dirty and may require *cleaning*; correcting errors, removing duplicates and resolving inconsistencies [Hernandez, et. al, 1998]

Name	Address	City	State	ZIP
Charles E Young Research Library	290 Charles E Young Dr N	LA	California	90095
Research Library	North 280 Charles Young Drive	Los Angeles	CA	
UCLA Charles Young Research Library	290 Charles Young Drive	Los Angelese		90059

This is especially challenging if data that identifies entities across sources is inconsistent; in this case, *entity resolution* is needed [Hernandez, et. al, 2013]

UC-Santa Barbara Branch

Title	Author	Dewey	Category
Fences	Wilson, August	812.54	Drama
The Hobbit	Tolkien, J. R. R.	823.912	Fiction
Team of Rivals	Goodwin, Doris Kearns	973.7092 B 22	Non-fiction

UCLA Library

Title	Author	ISBN
Fences	Wilson, August	0452264014
The Hobbit	Tolkien, J. R. R.	9780007440849
Team of Rivals	Goodwin, Doris Kearns	0743270754

UCLA Bookstore

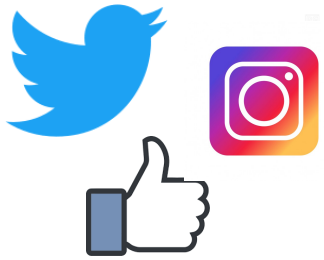
Title	ISBN	Price
Fences	9781557448040	12.06
Hobbit	8-0618002214	15.99
Team of Rivals	9780743270755	10.00

...And then Tim Berners-Lee brought us the world wide web, introducing a new generation of integration challenges.

Veracity: Data is no longer tightly controlled by DBAs and governed applications

RESTful API
GET PUT POST DELETE







Volume: There is a lot of it!

YouTube

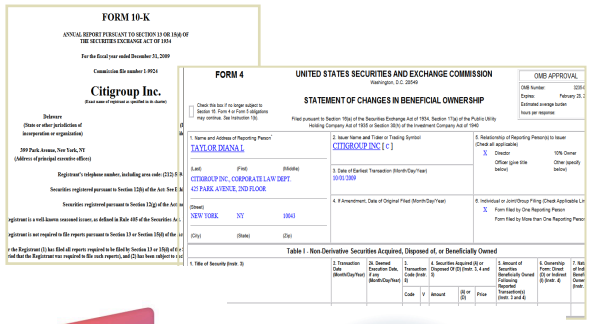
The Weather Company
An IBM Business

WIKIPEDIA
The Free Encyclopedia

DATA.GOV

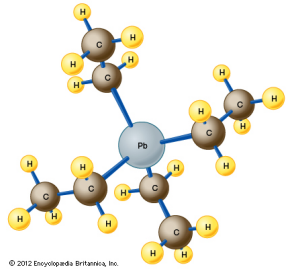
Velocity: It is being generated at a faster than we can consume!

Variety: Data is no longer structured











Relevant Publications

Relational and Distributed Databases

- E. F. Codd. 1970. A relational model of data for large shared data banks. *Commun. ACM* 13, 6 (June 1970), 377-387.
- M. M. Astrahan, M. W. Blasgen, D. D. Chamberlin, K. P. Eswaran, J. N. Gray, P. P. Griffiths, W. F. King, R. A. Lorie, P. R. McJones, J. W. Mehl, G. R. Putzolu, I. L. Traiger, B. W. Wade, and V. Watson. 1976. System R: relational approach to database management. *ACM Trans. Database Syst.* 1, 2 (June 1976), 97-137.
- M. Tamer Özsu, Patrick Valduriez: Distributed Database Systems: Where Are We Now? IEEE Computer 24(8): 68-78 (1991)

Data federation

- M. Tork Roth, M. Arya, L. Haas, M. Carey, W. Cody, R. Fagin, P. Schwarz, J. Thomas, and E. Wimmers. 1996. The Garlic project. In *Proceedings of the 1996 ACM SIGMOD international conference on Management of data* (SIGMOD '96), Jennifer Widom (Ed.). ACM, New York, NY, USA
- Alon Y. Levy, Anand Rajaraman, Joann J. Ordille: Querying Heterogeneous Information Sources Using Source Descriptions. VLDB 1996: 251-262
- Anthony Tomasic, Louiqa Raschid, Patrick Valduriez: Scaling Heterogeneous Databases and the Design of Disco. ICDCS 1996: 449-457
- Laura M. Haas, Donald Kossmann, Edward L. Wimmers, Jun Yang: Optimizing Queries Across Diverse Data Sources. VLDB 1997: 276-285
- Mary Tork Roth, Peter M. Schwarz: Don't Scrap It, Wrap It! A Wrapper Architecture for Legacy Data Sources. VLDB 1997: 266-275

Relevant Publications, cont'd

Data Warehousing

- William H. Inmon. 1992. *Building the Data Warehouse*. John Wiley & Sons, Inc., New York, NY, USA.
- Barry Devlin. 1996. *Data Warehouse: From Architecture to Implementation*. Lynne Doran Cote (Ed.). Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA
- Surajit Chaudhuri and Umeshwar Dayal. 1997. An overview of data warehousing and OLAP technology. *SIGMOD Rec.* 26, 1 (March 1997).

Schema Mapping

- Renée J. Miller , Laura M. Haas , Mauricio A. Hernández, Schema Mapping as Query Discovery, Proceedings of the 26th International Conference on Very Large Data Bases, p.77-88, September 10-14, 2000
- Sergey Melnik , Atul Adya , Philip A. Bernstein, Compiling mappings to bridge applications and databases, Proceedings of the 2007 ACM SIGMOD international conference on Management of data, June 11-14, 2007, Beijing, China
- Laura M. Haas, Mauricio A. Hernández, Howard Ho, Lucian Popa, and Mary Roth. 2005. Clio grows up: from research prototype to industrial tool. In *Proceedings of the 2005 ACM SIGMOD international conference on Management of data* (SIGMOD '05). ACM, New York, NY, USA, 805-810.

Relevant Publications, cont'd

Data Cleansing

- Mauricio A. Hernández, Salvatore J. Stolfo: Real-world Data is Dirty: Data Cleansing and The Merge/Purge Problem. Data Min. Knowl. Discov. 2(1): 9-37 (1998)
- Vijayshankar Raman, Joseph M. Hellerstein: Potter's Wheel: An Interactive Data Cleaning System. VLDB 2001: 381-390
- Erhard Rahm, Hong Hai Do: Data Cleaning: Problems and Current Approaches. IEEE Data Eng. Bull. 23(4): 3-13 (2000)

Entity Resolution

- Mauricio Hernández, Georgia Koutrika, Rajasekar Krishnamurthy, Lucian Popa, and Ryan Wisnesky. 2013. HIL: a high-level scripting language for entity integration. In *Proceedings of the 16th International Conference on Extending Database Technology* (EDBT '13). ACM, New York, NY, USA, 549-560.

New Challenges in Data Integration: Volume, Variety, Volume and Veracity

- Michael Franklin, Alon Halevy, and David Maier. 2005. From databases to dataspaces: a new abstraction for information management. *SIGMOD Rec.* 34, 4 (December 2005), 27-33.
- Bin He, Mitesh Patel, Zhen Zhang, and Kevin Chen-Chuan Chang. 2007. Accessing the deep web. *Commun. ACM* 50, 5 (May 2007), 94-101.
- Behzad Golshan, Alon Y. Halevy, George A. Mihaila, Wang-Chiew Tan: Data Integration: After the Teenage Years. PODS 2017: 101-106

