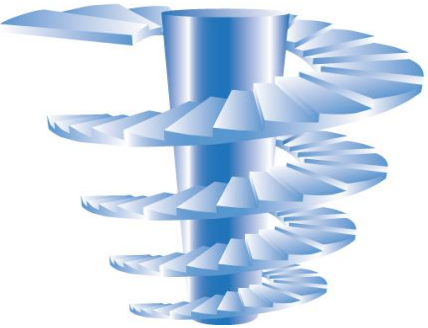


Synopsis of Big Data Technologies

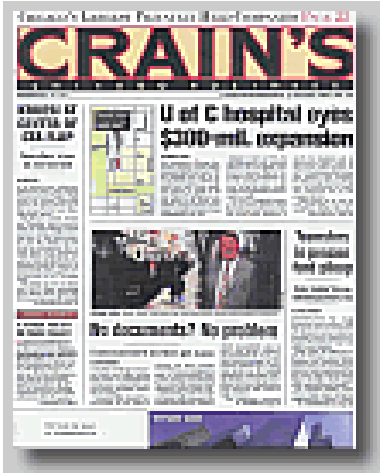
By David Marco

President

EWSolutions

EWSolutions' Background

EWSolutions is a Chicago-headquartered strategic partner and full life-cycle systems integrator providing both **award winning** strategic consulting and **full-service implementation services**. This combination affords our clients a full range of services for any size enterprise information management, meta data management, data governance and data warehouse/business intelligence initiative. Our notable client projects have been featured in the Chicago Tribune, Federal Computer Weekly, Journal of the American Medical Informatics Association (JAMIA), Crain's Chicago Business, and won the 2004 Intelligent Enterprise's RealWare award, 2007 Excellence in Information Integrity Award nomination and DM Review's 2005 World Class Solutions award.



Information Integrity Coalition

2007 Excellence in Information Integrity Award Nomination

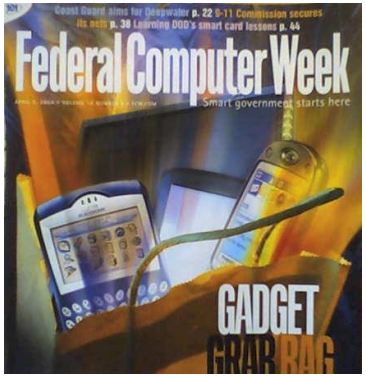
Chicago Tribune

intelligent enterprise REAL Awards TRANSFORM

2004 WINNER

Best Business Intelligence Application
Information Integration
Client: Department of Defense

JAMIA



DM Review 2005

World Class Solutions Award
Data Management

For more information on our Strategic Consulting Services, Implementation Services, or World-Class Training, email us at info@EWSolutions.com or call at 630.920.0005

EWSolutions' Partial Client List

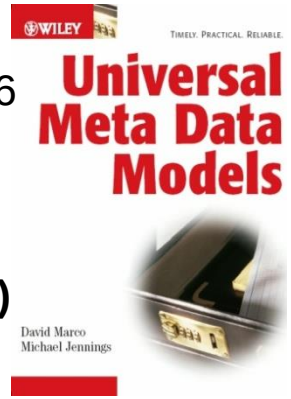
GSA Schedule

AFLAC
Arizona Supreme Court
Bank of Montreal
BankUnited
Basic American Foods
Becton, Dickinson and Company
Blue Cross Blue Shield companies
Booz Allen Hamilton
Branch Banking & Trust (BB&T)
British Petroleum (BP)
California DMV
California State Fund
Canadian National Railway
Capella University
Cigna
College Board
Comcast
Corning Cable Systems
Countrywide Financial
Defense Logistics Agency (DLA)
Delta Dental
Department of Defense (DoD)
Driehaus Capital Management
Eli Lilly and Company
Environment Protection Agency
Farmers Insurance Group
Federal Aviation Administration
Federal Bureau of Investigation (FBI)
Fidelity Information Services
Ford Motor Company
GlaxoSmithKline
Harbor Funds
Harris Bank
The Hartford
Harvard Pilgrim HealthCare
Health Care Services Corporation
Hewitt Associates
HP (Hewlett-Packard)
Information Resources Inc.
International Paper
Janus Mutual Funds
Johnson Controls
Key Bank
LiquidNet
Loyola Medical Center
Manulife Financial
Mayo Clinic
McDonalds
Microsoft
MoneyGram
NASA
National City Bank
Nationwide
Neighborhood Health Plan
NORC
Physicians Mutual Insurance
Pillsbury
Quintiles
Sallie Mae
Schneider National
Secretary of Defense/Logistics
Singapore Defense Science & Technology Agency
Social Security Administration
South Orange County Community College
Standard Bank of South Africa
SunTrust Bank
Target Corporation
The Regence Group
Thomson Multimedia (RCA)
Thrivent Financial
United Health Group
United Nations (ICAO)
United States Air Force
United States Army
United States Department of State
United States Navy
United States Transportation Command
University of Michigan
University of Wisconsin Health
USAA
US Cellular
Waste Management
Wells Fargo
Wisconsin Department of Transportation
Zurich Cantonal Bank

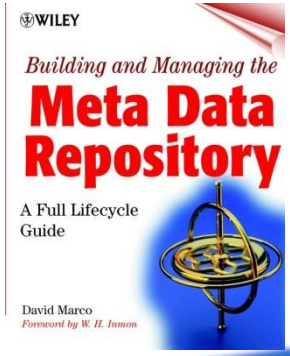
For more information on our Strategic Consulting Services, Implementation Services, or World-Class Training email us at **Info@EWSolutions.com**

David Marco – Professional Profile

Best known as the world's foremost authority on meta data management and the father of the Managed Meta Data Environment, he is an internationally recognized expert in the fields of data governance, big data, data warehousing, master data management and enterprise information management (EIM). In 2004 David Marco was named the **"Melvil Dewey of Metadata"** by **Crain's Chicago Business** as he was selected to their very prestigious **"Top 40 Under 40"** list. David Marco has authored several books including the widely acclaimed **"Universal Meta Data Models"** (Wiley, 2004) and the classic **"Building and Managing the Meta Data Repository: A Full Life-Cycle Guide"** (Wiley, 2000).



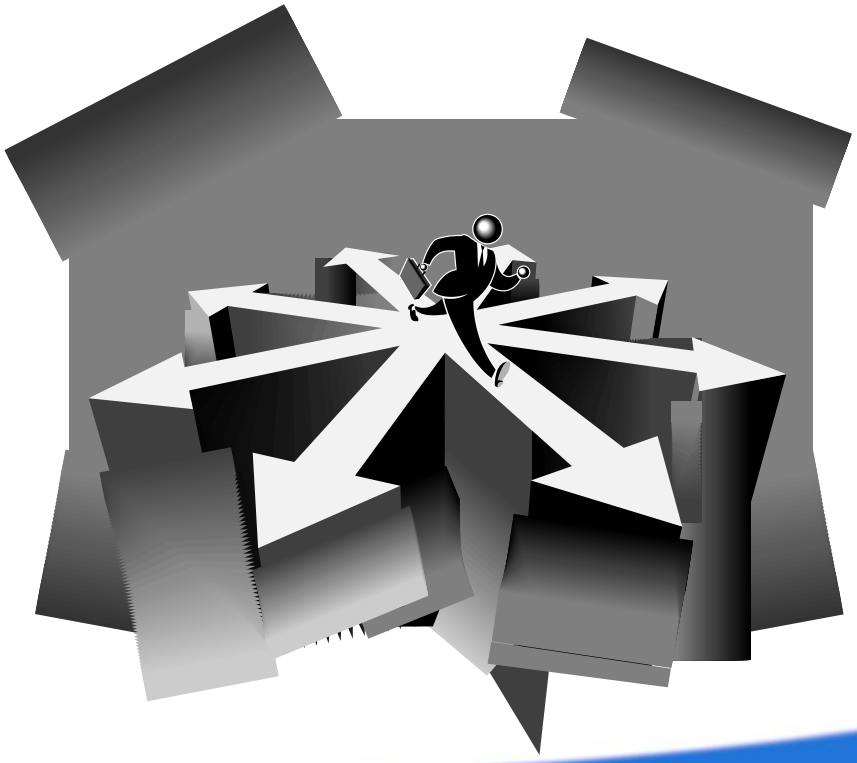
- **2014** EWSolutions was inducted into the Hinsdale **business Hall-of-Fame** after 6 consecutive years of receiving "Best of" awards in Enterprise Information Management
- Selected to the prestigious **2004 Crain's Chicago Business "Top 40 Under 40"**
- **2008 DAMA Data Management Hall of Fame (**Professional Achievement Award**)**
- **2007 DePaul University named** him one of their **"Top 14 Alumni Under 40"**
- Presented hundreds of keynotes/seminars across four continents
- Published hundreds of articles on information technology
- Author of several best selling information technology books
- Taught at the **University of Chicago** and **DePaul University**
- Holds both a CDMP and a CBIP certification



Email: DMarco@EWSolutions.com

Agenda

- Synopsis of Key Big Data Technologies
- 5 V's of Big Data
  - Volume
  - Variety
  - Veracity
  - Velocity
  - Value



Key Definitions

Hadoop

- Apache Hadoop is an open-source software framework that supports data-intensive distributed applications
- Hadoop was derived from Google's GFS (Google File System) and MapReduce systems
- The library, instead of the hardware, is designed to detect and handle failures, to deliver highly-available service on top of a cluster of computers
  - **Google File System (GFS or GoogleFS):** is a proprietary distributed file system developed by Google for its own use. It is designed to provide efficient, reliable access to data using large clusters of commodity hardware
  - **Hadoop Common:** The common utilities that support the other Hadoop modules.
  - **Hadoop Distributed File System (HDFS™)**: A distributed file system that provides high-throughput access to application data
  - **Hadoop Yet Another Resource Negotiator (YARN)**: A framework for job scheduling and cluster resource management
  - **Hadoop MapReduce**: A YARN-based system for parallel processing of large data sets

* *Some sections of this presentation adapted from Dr. Anne Marie Smith's presentation on Big Data at 2015 Enterprise Data World*

Purpose of Hadoop

- Hadoop implements Google's MapReduce, using HDFS
- MapReduce divides applications into many small blocks of work
- MapReduce can then process the data where it is located
- Hadoop's processes can run on thousands of clusters, making processing very fast

Why is Hadoop Different?

- **Scalable:** Reliably store and process high volumes of data (e.g. petabytes)
- **Economical**: Distributes the data and processing across clusters of commonly available computers (in thousands)
- **Efficient:** By distributing the data, it can process data in parallel on the nodes where the data is located
- **Reliable:** Automatically maintains multiple copies of data and automatically redeploys computing tasks based on failures
- Anybody notice anything in the **"Reliable"** bullet?

NoSQL

- NoSQL (Not Only SQL) is a database format that provides for storage and retrieval of data that is modeled in ways other than tabular formats used in relational databases
- NoSQL databases give more flexibility and availability for less-structured data, and can make some operations faster, especially when traversing very large unstructured datasets

Types of NoSQL Databases

- The list of NoSQL database types seems to change on a daily basis
- We will discuss the following 4 types NoSQL databases:
  - **Key-Value**
  - **Document**
  - **Column-Oriented**
  - **Graph**

Types of NoSQL Databases

- **Key-Value:** have a single table with key-value pairs, meaning two columns: one being the (Primary) Key, and the other being the Value
  - generally useful for storing session information, user profiles, preferences, shopping cart data, unstructured data
- **Document:** these types of NoSQL differ from each other but that are built to store documents and encode data (or information) in some standard format(s)
  - generally useful for content management systems, blogging platforms, web analytics, real-time analytics, ecommerce-applications

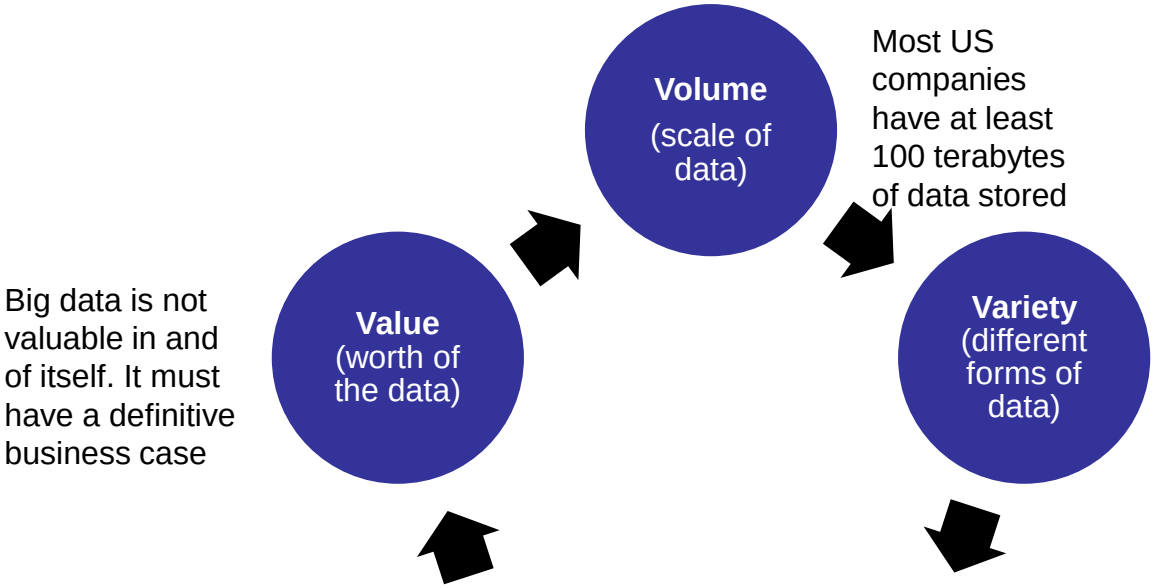
Types of NoSQL Databases

- **Column-Oriented:** each storage block contains data from only one column. RDBMS store a single row as a continuous disk entry. Different rows are stored in different places on disk while Columnar databases store all the cells corresponding to a column as a continuous disk entry thus makes the search/access faster
  - generally useful for content management systems, blogging platforms, maintaining counters, expiring usage, heavy write volume such as log aggregation

- **Graph:** are the kind of NoSQL database designed for data whose relations are well represented as a graph (elements interconnected with an undetermined number of relations between them)
  - very well suited to problem spaces for connected data, such as social networks, spatial data, routing information for goods, social relations, road maps, network topologies, money, recommendation engines

Demystifying Big Data Technologies

5 V's of Big Data



Most US companies have at least 100 terabytes of data stored

4+ billion hours of content are watched monthly on YouTube alone

Poor data quality cost the US economy $3.1 trillion in 2012 (Gartner)

Every session of the NYSE captures at least 1 TB of data

Big data is not valuable in and of itself. It must have a definitive business case

Questions