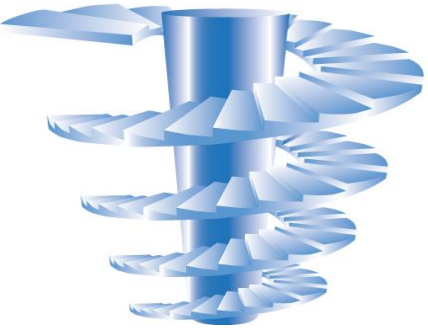


# Big Data Meets Data Management

**By David Marco**

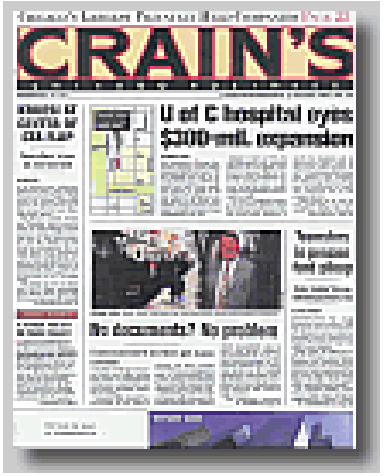
*President*

***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*

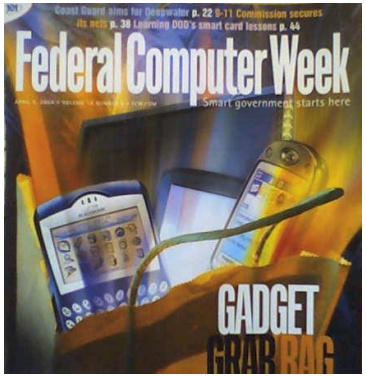
Chicago Tribune

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

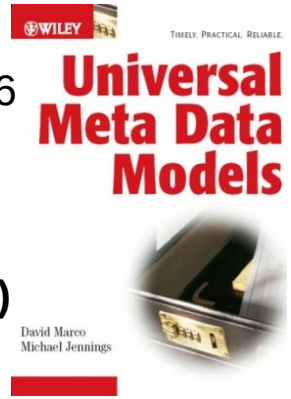
GSA Schedule

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

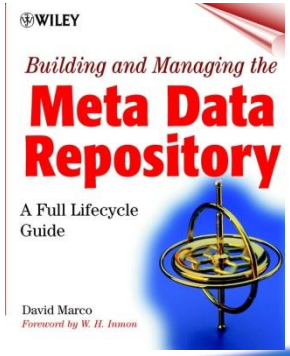
For more information on our Strategic Consulting Services, Implementation Services, or World-Class Training email us at **Info@EWSolutions.com**

# 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).



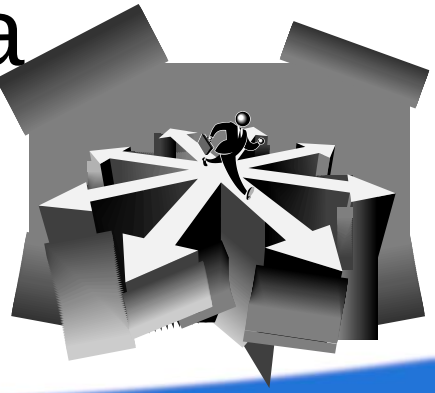

- **2014** EWSolutions was inducted into the Hinsdale **business Hall-of-Fame** after 6 consecutive years of receiving "Best of" awards in Enterprise Information Management
- Selected to the prestigious **2004 Crain's Chicago Business "Top 40 Under 40"**
- **2008 DAMA Data Management Hall of Fame (**Professional 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

- Big Data Technology Wrap Up
  - Schema-less Data
  - Big Data Technology by Data Management Function
- Big Data Growth
- Is Big Data Successful?
- Big Data Top 10 List
- Data Management and Big Data



# Schema vs. Schema-Less Data

# Schema vs. Schema-less

- All NoSQL databases claim to be schema-less
- Schema-less means there is no schema enforced by the database itself
- **Rather the scheme is built into the access code**
- Databases with strong schemas, such as relational databases, can be migrated by saving each schema change, plus its data migration, in a version-controlled sequence
- Schema-less databases need careful migration and data management due to the implicit schema in any code that accesses the data

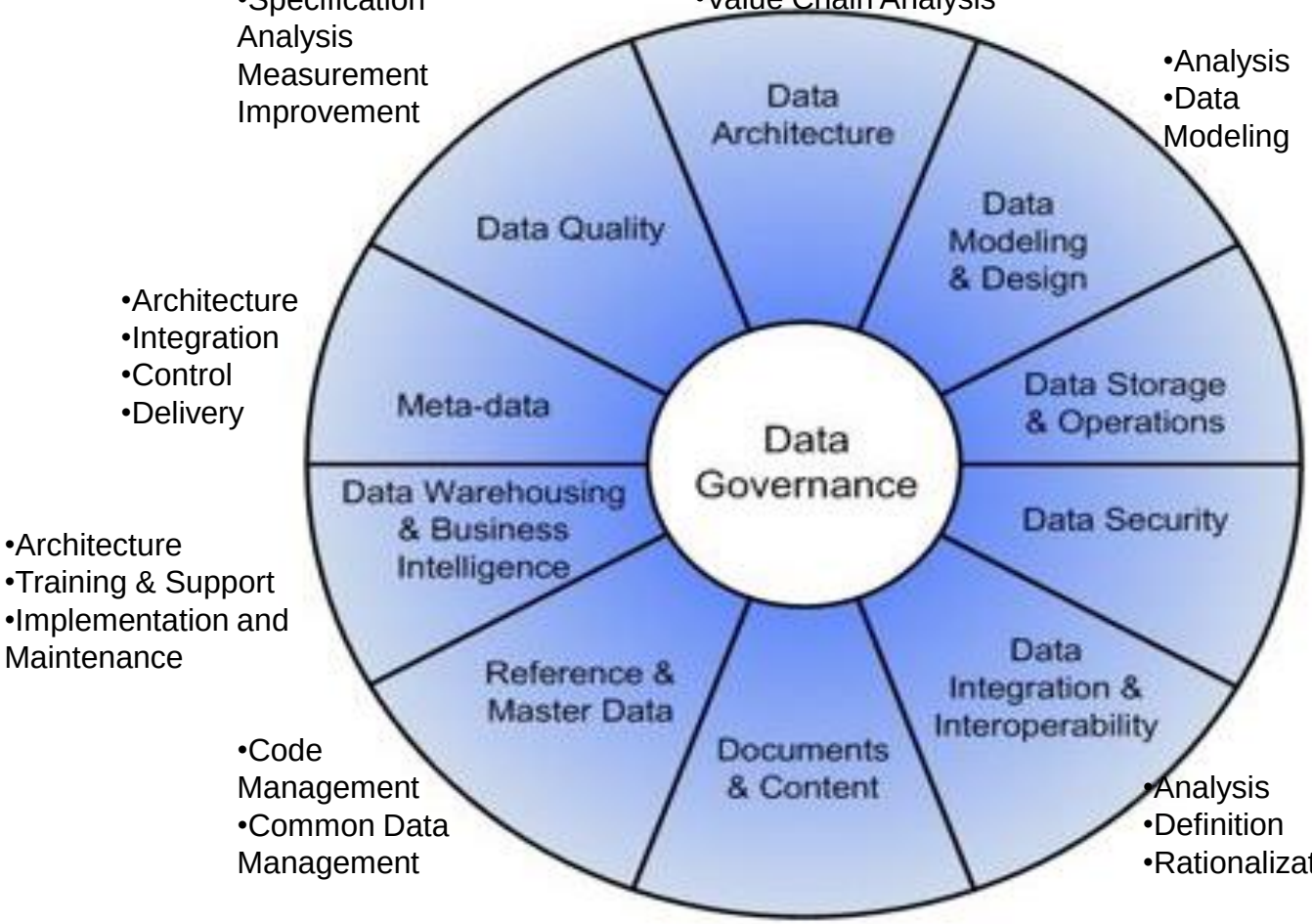
* *Some sections of this presentation adapted from Dr. Anne Marie Smith's presentation on Big Data at 2015 Enterprise Data World*

# Big Data Technologies Applied to DAMA-DMBOK Functions

# DAMA-DMBOK ©



# Data Storage and Operations

The technologies and processes organizations use to maximize or improve the performance of their data storage resources


*File system that provides the ability to store large volumes of structured and unstructured data*


*Operations, resource (node), and scheduling management for write and read to the cluster*


*Workflow scheduling component for data transformations*


*Manages services, configurations, and their synchronization across the cluster*

# Data Integration

The combination of technical and business processes used to combine data from disparate sources into meaningful and unified view, according to business requirements and accepted practices

Sqoop

*Provides ability to import data from a RDBMS to HDFS.*


*Provides ability to collect, aggregate, and move huge log files ). into HDFS (e.g., apps, GPS, social, sensors, other).*

Apache Kafka

*Provides high volume fault tolerant publish & subscribe messaging for real-time analysis.*

APACHE STORM™

*Provides real-time processing of data streams for monitoring and alerts.*

# Data Quality

A measure of the degree to which data satisfies the information needs of its consumers, reflects the nature and state of the real world concepts to which it relates, is coherent within itself, and provides value in the decision-making processes for which it is to be utilized


*Provides relational structure to HDFS data. File formats can be applied to data from HDFS or local file system*


*Provides ability to import data from a RDBMS to HDFS. Imported data can be constrained through import control arguments and basic SQL execution.*


*Provides ability to collect, aggregate, and move huge log files ). into HDFS (e.g., apps, GPS, social, sensors, other). Flume agent can be use with predefined data patterns (sinks) to ensure data format.*

# Meta Data Management

All the physical data and knowledge about the business and technical processes used by an organization.  Meta data is knowledge about the organization's data

Falcon

*Provides data lineage between data sources and the  cluster including integration with the metastore/catalog (e.g., Hive HCatalog).*

Hive/HCatalog

*Provides relational structure to HDFS data. File formats can be applied  to data from HDFS or local filesystem*

# Documents and Content

The management of documents and non-structured content found in audio, video, email, images, etc. and the meta data associated with this material


*Provides ability to collect, aggregate, and move huge log files ). into HDFS (e.g., apps, GPS, social, sensors, email, other).*


*Provides ability to search of data in the cluster by indexing to enable full text search.*

# Data Warehouse and BI

A data warehouse is a subject-oriented, integrated, time-variant and non-volatile collection of data in support of management's decision making process. Business Intelligence (BI) is the collection of activities that allow an organization to analyze data and make decisions based on facts from historical and predictive data sets


*Provides compute algorithm typically used to produce output data from a large volume of data in the cluster for consumption.*


*Provides a enhanced compute approach typically used to produce output data from a large volume of data in the cluster for consumption.*


*Provides a in-memory compute method typically used to produce output data from a large volume of data in the cluster for consumption (e.g., machine learning algorithms).*


*Provides fast big table access to large quantities of data typically on top of the cluster.*


*Provides semantic layer for accessing data in the cluster.*

# Data Security

Data security concerns the protection of data from accidental or intentional but unauthorized modification, destruction or disclosure through the use of physical security, administrative controls, logical controls, and other safeguards to limit accessibility

KNOX

*Provides service level authorization for users/groups.*

Ranger

*Provides security authorization (grant/revoke), policy administration, and audit for the cluster.*

Hive/HCatalog

*Provides semantic layer (table) for accessing data in the cluster that can be secured.*

# Data Governance

Provides the organizational oversight, processes and methods to effectively manage data as an asset across the organization


Hive/HCatalog

*Provides relational structure to HDFS data. File formats can be applied to data from HDFS or local filesystem*

Falcon

*Provides data lineage between data sources and the cluster including integration with the metastore/catalog (e.g., Hive HCatalog).*


Ranger

*Provides security authorization (grant/revoke), policy administration, and audit for the cluster.*


*Provides ability to search of data in the cluster by indexing to enable full text search.*

# Master and Reference Data

Data about core business entities and concepts, independent of transactions, and data that defines the set of permissible values to be used by other data fields

Sqoop

*Provides ability to import data from a RDBMS to HDFS.*

Hive/HCatalog

*Provides semantic layer for accessing data in the cluster.*

# Big Data Growth

# Explosive Data Growth

Every minute of the day:

- YouTube receives 48 hours of uploaded video
- Over 2 million search queries hit Google
- Twitter users post about 100,000 tweets
- 571 new websites are created
- Over 200,000,000 email messages are created and sent

http://ftp-hosting-services-review.toptenreviews.com/the-explosive-growth-of-digital-data.html

# Need for Data Management

- Trends in data growth show need for data management in every industry (Gartner 2014, US Bureau of Labor Statistics 2014, ComputerWorld 2014, etc..)
  - Data Architect
  - Data Modeler
  - Data Governance specialist
  - Data Quality specialist
  - Master Data Management specialist
  - Etc……

# Is Big Data Successful?

# Big Data Success

- In 2013 companies spent $31 billion on Big Data
- The market is expected to top $114 billion by 2018
- 60% of executive believe Big Data will upend their industries within 3 years*
- 55% of Big Data projects do not get completed**
- Through 2017, 60% of Big Data projects will fail to go beyond piloting and experimentation and will be abandoned

* ***"Cracking the Data Conundrum: How Successful Companies Make Big Data Operational",*** *Cap Gemini, January 14, 2015*

** ***"CIOs & Big Data : What Your IT Team Wants You To Know",*** *infochimps, 2013*

*** *Ted Friedman, VP of Gartner, Twitter 12/29/2014*

# Big Data Success

- Executives describe the own Big Data projects as*:
  - 8% “Very Successful”
  - 27% “Successful”
  - 65% Below “Successful”
- Some other sources show Big Data failure rates at 30%, 50%, 75% and one showed 90%**

* ***“Cracking the Data Conundrum: How Successful Companies Make Big Data Operational”,*** *Cap Gemini, January 14, 2015*

** ***“Marketing IT In-House: Don’t Let BI Failure Statistics Stop You”,*** *Max T. Russell, February 18, 2014*

# Top 10 Reasons for Big Data Failure

# #1: Believing the Hype

- Big Data have been over hyped and over marketed, especially to the C-Level
- They believe that it is a magic bullet
- In reality, most don't even know what Big Data actually means
- They buy into the marketing and the false promises
- They attempt overly ambitus projects

# #2: Using Big Data When There Wasn't a Need

- Remember our definition?
- **Big Data:** a collection of data sets so large and complex that it becomes difficult to process using traditional database management tools or conventional data processing applications
- Big Data should not be used when a traditional database will suffice

# #3: Big Data Silos

- Do you know what a Data Lake is?
- **Data Lake:** is a large data repository sourced from multiple applications where the data is stored in its native format
- Most Big Data initiatives are stand-alone efforts and lack integration
- They are just silos of data; regardless, of if they are Big or not

# #4: Poor Business Objectives

- Many Big Data projects start with poor or unclear business objectives
- Big Data is **NOT** science projects!!
- They should have a clear, definable and profitable business case

# #5: Lacking the Right Skills

- Big Data skills are sparse in the industry
- Truly qualified people are difficult to find and are likely to be hired away

# #6: Management Failure

- Many companies select the wrong uses for Big Data
- Most management users their “gut” to make decisions as opposed to trusting the actual analytics being presented

# #7: Big Data Waterfall

- Too many Big Data projects are ran as projects and not programs
- They use a Waterfall methodology, instead of an iterative approach
- Big Data programs are best built iteratively and over time

# #8: Targeting Many Petabytes

- Big Data technology can handle petabytes of data and over time it will have greater capacity
- In general, going beyond 1 – 3 petabytes is very dangerous and takes high levels of expertise and money

# #9: Can You Feel The Heat?

- Newer server cabinets are much more powerful than before and the also produce much more heat
- Temperatures in most data centers aisles range from 80 to 115 degrees
- Larger data centers can be 7 – 9 times the size of a football field
- Typically a 1/3 of the energy consumed by a Big Data center is used by the cooling system for the facility

# #10: Lack of Data Management

- Common complaints heard during Big Data failures include:
  - Poor data quality
  - Business didn't understand the analytics they were being presented
  - Ineffective governance models
  - Data security and privacy issues
- This is all data management!!

# Data Management Meets Big Data

# Big Data Goals

- Better decisions through unstructured analytical data & traditional structured data
- Allow the organization to manage extremely large sets of structured and unstructured data for operations and analytical functions

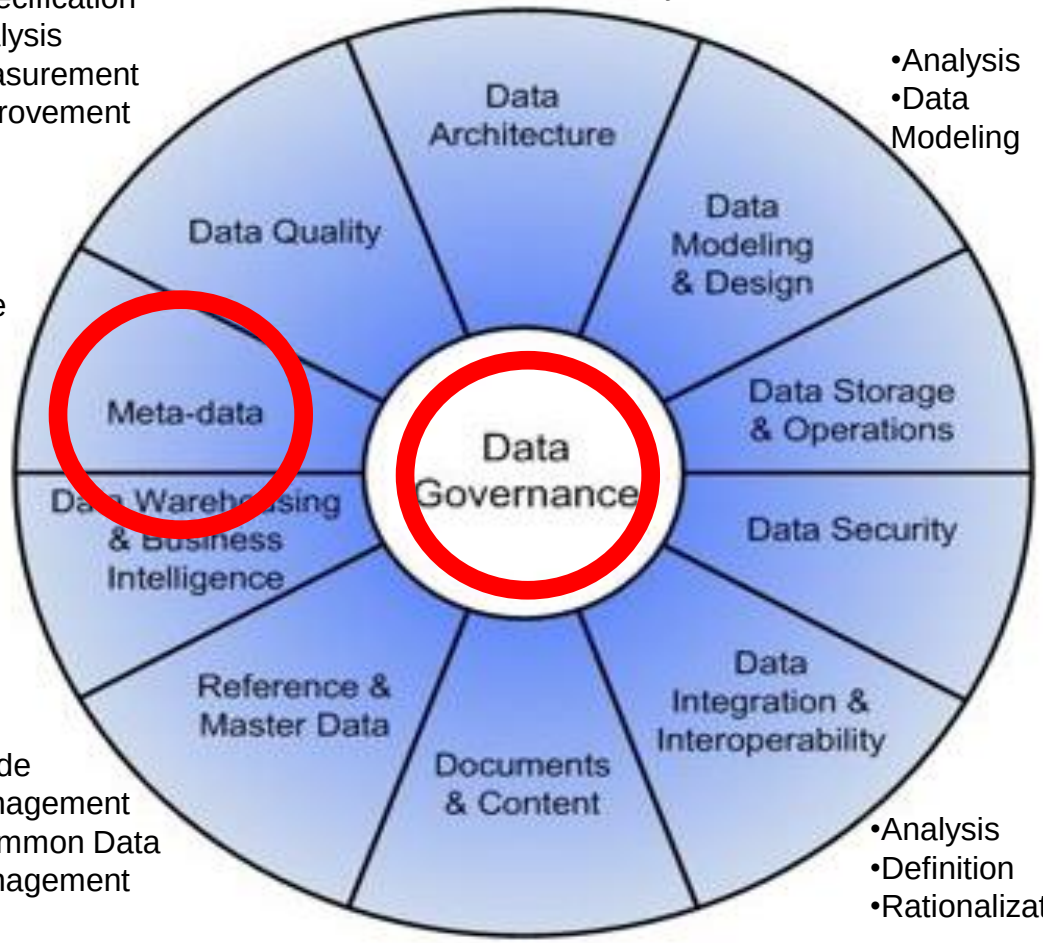

 

# Big Data Meets Data Management

- Many companies start by building a proof of concept
- Early projects might work well enough because they are constrained (one or a few large data sources and are not integrated)
- Initial success seems to imply that it is safe to ignore data management principles like meta data management, master data management and data governance
- **This is a massive mistake!!**
- The key word in “Big Data” is DATA and it must be managed
- Meta Data Management, Data Governance and all of our Data Management disciplines become highly magnified in Big Data implementations as the stakes are higher (data volumes, costs, etc., etc) and we lose the RDBMS

# DAMA-DMBOK ©



# Data Governance and Big Data

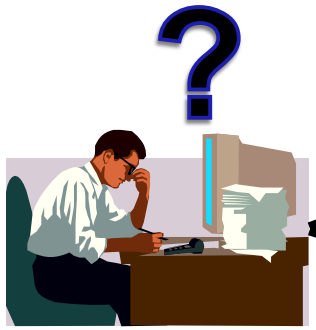
# Data Governance Defined

- ***Data Governance:*** defines the people, processes, framework and organization necessary to ensure that an organization's information assets (data and meta data) are formally, properly, proactively and efficiently managed throughout the enterprise to secure its trust, accountability, meaning and accuracy

# Understanding Data Governance

## DATA GOVERNANCE



I need to make profitable decisions

I don't know what I'm looking at

**Data**

Misunderstood
Inaccurate
Misleading

*What Do I Do?*

Data Stewards

**Data Governance**

*Policies, Procedures, Consensus, Knowledge, Information, Data, Meta Data*

Transform Data Into Information

Understood
Accurate
Consistent

Data Stewards

**Actionable Information**

# Data Governance and Big Data

- *"Enterprise Data Management and Big Data Management require the foundation of leadership that is generated from a strong, stable enterprise data governance program that includes both traditional and big data. Effective data governance is essential to the success of any organization as they move to more analysis and faster operations."*

*Gartner, April 24, 2013.*

# Data Governance Policies

- Data Governance Policies and Standards are aligned and processes integrated with:
  - Comprehensive Information Security Policies and Standards
  - HIPAA / Privacy Policies
  - Applicable Legal Policies
  - Risk Management Guidelines
  - Vendor Collaboration
  - Business Area Policies, Standards, Guidelines, and Procedures

*How does inclusion of Big Data affect each policy, standard and its governance?*

# Meta Data and Big Data

# Meta Data Fundamentals - 1

## *Business Metadata Example*

- Business metadata is metadata about the business terms, business processes and business rules.
- Business metadata provides the semantic layer between your systems and their business users.
- It provides users a roadmap for navigating all the data in the enterprise by documenting what information is available and, when accessed, provides a context for interpreting the data.

**Invaluable for making sound business decisions.**

| Business Metadata | Definition | Standards | Examples |
|---|---|---|---|
| Entity Business Name | Contains the common name of the Entity that is recognized by business users. | ISO/IEC 11179-5 Information Technology – Metadata Registries Part 5: Naming and identification principles. | Customer (versus technical name of CUST_NM) |
| Entity Business Description | Contains the detailed explanation of the business meaning of the Entity in the context of the enterprise. | ISO/IEC 11179-4 Information Technology – Metadata Registries Part 4: Formulation of data definitions | e.g., Customer definition: A current or potential user/buyer of products or services from ACME. |
| Data Attribute Business Acronym | Common, business recognized acronym coding of the data attribute (If applicable) | Industry standards (e.g., ISO, Industry), ACME, or application specific | e.g., SSN (Social Security Number) |
| Data Attribute Business Description | Detailed explanations of the business meaning of the data attribute. | ISO/IEC 11179-4 Information Technology – Metadata Registries Part 4: Formulation of data definitions | e.g., Customer Name definition: The legal name of a current or potential user/buyer of products or services from ACME. |
| Data Attribute Business Name | Contains the common name of the data attribute that is recognized by business users. | ISO/IEC 11179-5 Information Technology – Metadata Registries Part 5: Naming and identification principles. | e.g., Customer Name |

# Big Data needs meta data!

# Meta Data Fundamentals - 2

## *Technical Metadata Example*

- Technical metadata is metadata describing technical aspects of IT systems, which designers and developers use to build and maintain them.
- Examples of technical metadata include descriptions of database tables, data attributes, sizes, data types, database key attributes and indices and technical data transformation rules.

| Technical Metadata | Definition | Standards | Examples |
|---|---|---|---|
| Data Attribute Data Type | Data type associated with the data attribute | | Date, Number, String |
| Data Attribute Physical Name | Represents the physical data attribute name, DBMS or data grouping, of the data attribute. | ACME, IT - Information Management Policies and Procedures, Data Modeling Standards | e.g., CUST_NM (for Customer Name) |
| Data Entity Type | Describes what data format the database is represented in. | | e.g., XML, Relational, Message, or Flat File, other |
| Database Physical Name | The physical name of a database or data package | ACME, IT - Information Management Policies and Procedures, Data Modeling Standards | e.g., PCSDB02 (LDB), PANDDB01 (IC+), PCSDB01 (Kronos) |
| Derived Data Attribute Flag | Denotes a data attribute instance whose value is obtained from some expression or business rule dependency on other data attributes. | | "Y" for derived attribute |
| Source Application | The recognized name of the source system based on the ACME business case name assigned to the project | ACME IT and/or PMO Application naming conventions | e.g., LDB, Intercom Plus (IC+), Kronos |
| Entity Physical Name | The physical name of a data source table or data grouping | ACME, IT - Information Management Policies and Procedures, Data Modeling Standards | e.g., Examples: sum_total_accounting total_sales |
| Entity Refresh Type Name | Indicates the frequency with which the table content is refreshed. | | e.g., Intraday, Daily, Weekly, Monthly, etc. |
| Data Attribute Business Rule Description | An explanation of the criteria and constraints that apply to population of the data attribute. | | |
| Data Attribute Business Rule Name | A unique name for a data attribute business rule | | |

Big Data needs meta data!

# What is Meta Data for Big Data?

- Most organizations collect Big Data for analytics and sharing
- Cannot perform analytics or share data effectively without the use of business and technical meta data
- Organizations that want to use big data effectively are building strong meta data capabilities, along with data governance, to catalog technical meta data and record business meta data, then to enrich that catalog with additional meta data from the results of the analytics and sharing activities
  - Need a managed meta data environment (MME)

# Managed Meta Data Environment

## Meta Data Sourcing Layer

Software Tools

Meta Data Extract

Documents/ Spreadsheets

Meta Data Extract

Messaging/Transactions (EAI, web services, XML, etc.)

Meta Data Extract

Applications (CRM, ERP, data warehouses, etc.)

Meta Data Extract

Websites/ E-Commerce

Meta Data Extract

Third Parties (business partners, vendors, customers, government agencies)

Meta Data Extract

Meta Data Integration Layer

Meta Data Management Layer

Meta Data Repository

End Users (business and technical)

## Meta Data Delivery Layer

Websites/E-Commerce

Third Parties (vendors, customers, government agencies)

Messaging/Transactions (EAI, web services, XML, etc.)

Business Users

Data Warehouse/ Data Mart(s)

End Users (business and technical)

Meta Data Marts

End Users (business and technical)

Applications (CRM, ERP, etc.)

Software Tools

# In Conclusion

- Big Data should only be used when the volumes exceed traditional database technologies
- There are **NO MAGIC BULLETS**
- You must ALWAYS understand your data (data management)
- Get ready for growing pains
- Be smart, diligent and patient
- You will achieve **GREAT** results

# Questions