Enterprise Architecture led Data Quality Strategy

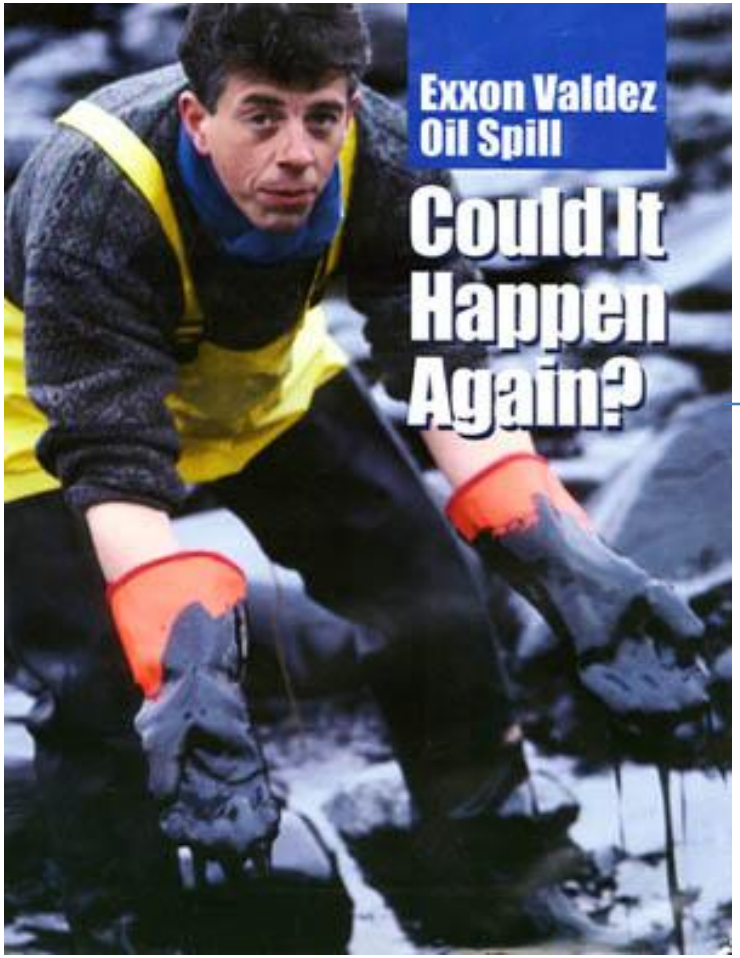
"Two Sides of the Same Coin"

MIT IQIS, July 14, 2010

Jay Barua

Corporations annually lose billions of dollars to clean up bad data. Like having one major oil spill every year. Imagine.

1989



Yes it can

2010

Losses due to bad data unimaginable

- Study by AT Kearney shows
  - Retail companies lose over $40 billion every year due to supply chain inefficiencies
  - 30 % of item data in catalogs is in error and each item cost on the average $75 to clean
  - 60 % of all invoices have errors and it takes approx $220 average to reconcile each single invoice
  - Errors happen because data like item, price, units of sale etc is out of sync between businesses doing B2B commerce
  - Lack of data governance contributes largest to the above problem

Market Demands Clean Data

- Recent upsurge in Retail online sales highlights need for clean data

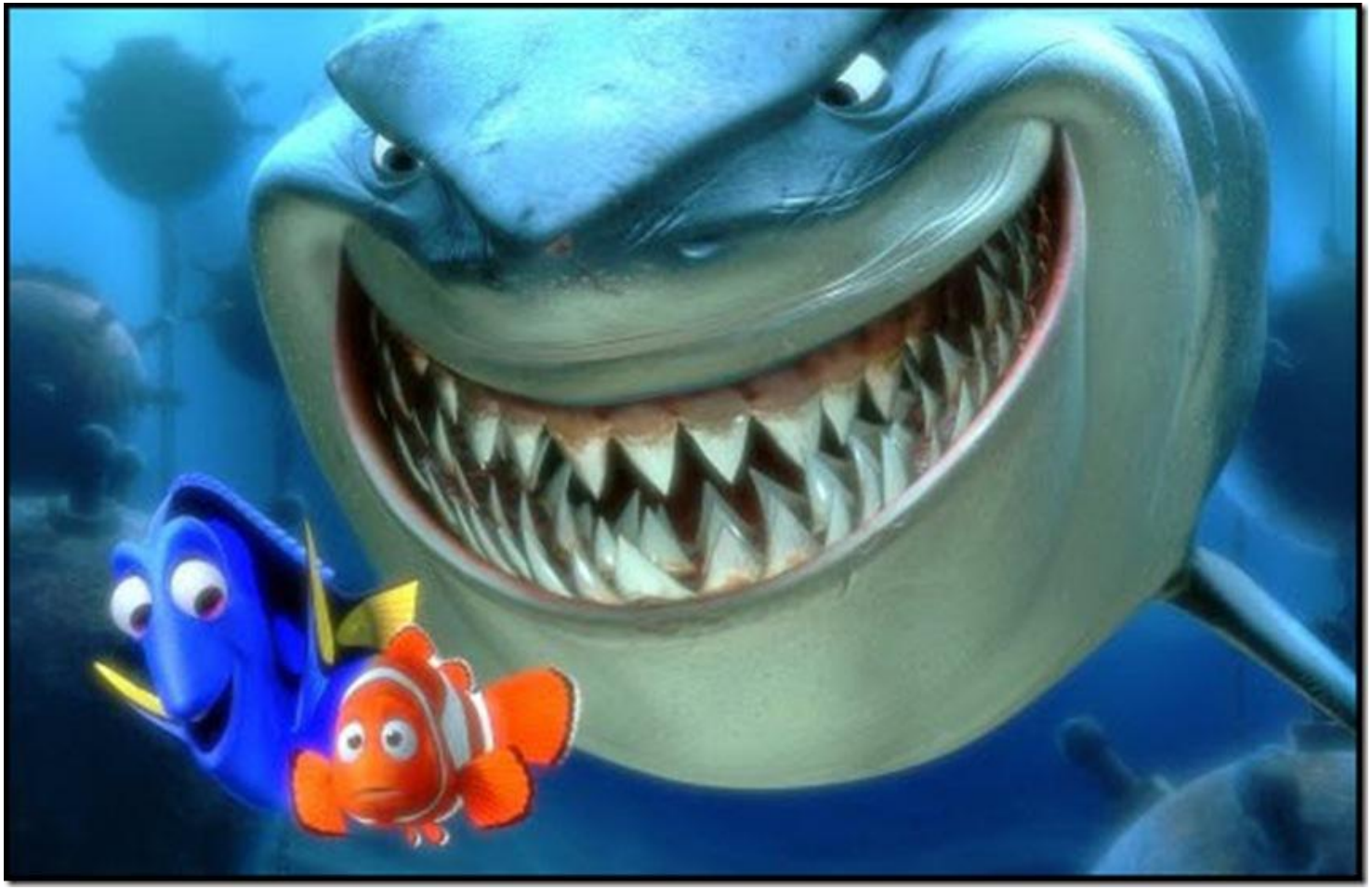
  - Online retailers spend millions of dollars to clean inaccurate information available at checkout process

  - Overstock.com spent over $200,000 a year correcting bad addresses that had negatively impacted shipment delivery and raised costs and above all made customers unhappy

  - Overstock.com saved more than $1M every year once they adopted automatic address verification system

Data Quality

So Let us find Data Quality before IT finds you (Courtesy Finding Nemo by Pixar and Jim Harris)



What is Data Quality?

- Wikipedia defines Data are of high quality "*if they are fit for their intended uses in operations, decision making, and planning* " (J.M Juran)

  - Some people view data quantitatively
  - Others look at it qualitatively

Quantitative Data

Put a number and correct it !

- How many customer records in my data warehouse have more than one primary email?

- How many times in a year have we sent wrong emails to wrong customers ?

Qualitative Data

- Data that is Bad for one can be Good for other

  - Customer and email matching processes in a system that takes online orders can be different on a system that tracks gift card sales for the same customer

  - Same customer may have different emails in both these systems

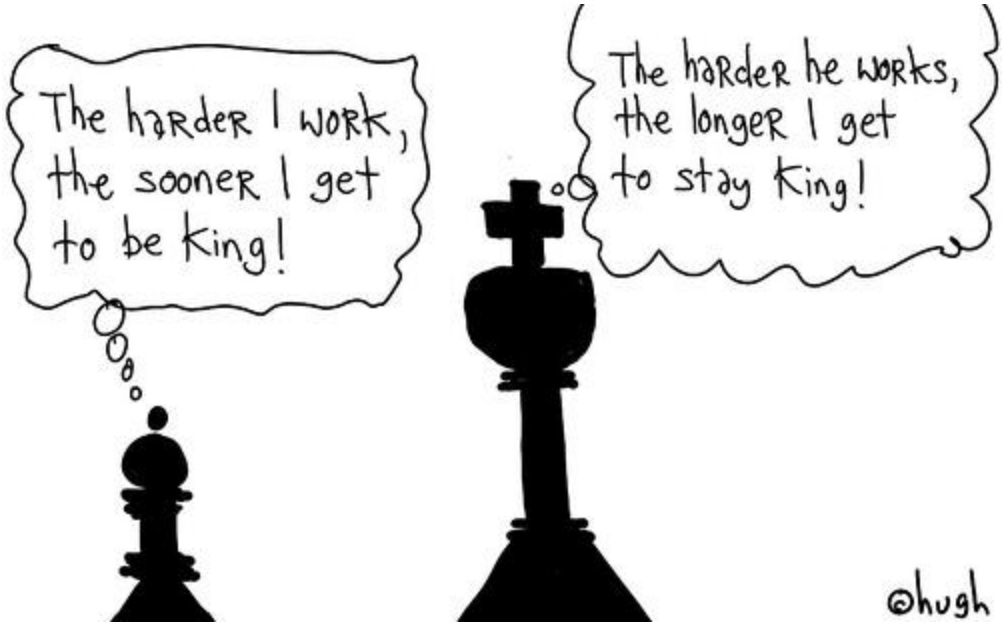
Why does data go bad ?

- Unmanageable **Silos** creates bad data
  - Different versions of data can create different interpretations
- Disconnected **Processes** creates bad data
  - Disconnected process acting on a data supply chain can make the data bad
- Lack of **Governance** creates bad data
  - Lack of Governance can create disconnected data

Can Enterprise led processes help ?

Enterprise Architecture

From Incite comes Insight (Courtesy Terrybone)



Let us Create a Data Quality Lab ?

By Investing in Enterprise Architecture (EA)

- Adopt a Data Quality Strategy that works for you
- Set up a simple Enterprise Data Architecture framework

Build your Data Quality (DQ) strategy with

- Data Governance
- Data Integration
- Metadata Management

Use Data Governance with ***tactical policing***

- Acquire management responsibility
- Appoint and empower data quality Inspectors
- Set standard data definitions
- Follow culturally accepted evolutionary process
- Use end-end benchmarked processes

Data Integration promotes rich data

- Evaluate & migrate core legacy systems to common platform
- Create comprehensive view of key data – Customers & Products
- Extended your workflow
- US Xpress saves $6 Million a year using data integration methods to clean location information and manage truck idle time. They converted their “bad” data to “rich” data

Data Production Factory (DPF)

	PRODUCT	DATA
INPUT	Raw Materials	Raw Data
PROCESS	Processing	Processing
OUTPUT	Physical Product	Data Product

Analogy between Physical and Data Products (Wang et al [5])

Metadata Management

- Let us have a name for everything !
  - Important to define data so that correct decisions are taken which in turn produces more clean data
  - Important to define data ancestry so that we can explain the how data is transformed in a data production factory
  - Effective schema attribution provides meaning to data

Seven DQ Enablers

- Completeness – Is all information available ?
- Conformity – Agrees on certain formats
- Consistency – Does data instantiation provide same value
- De-duplication – Try to maintain one version of truth
- Uniqueness – Does collection of data maintain identity ?
- Integrity – Dependencies between entities maintained
- Accuracy – Real world representation

Create your EA framework to enable DQ

- Data Profiling
- Data Auditing
- Exception Management
- Data Integration
- Data Architecture

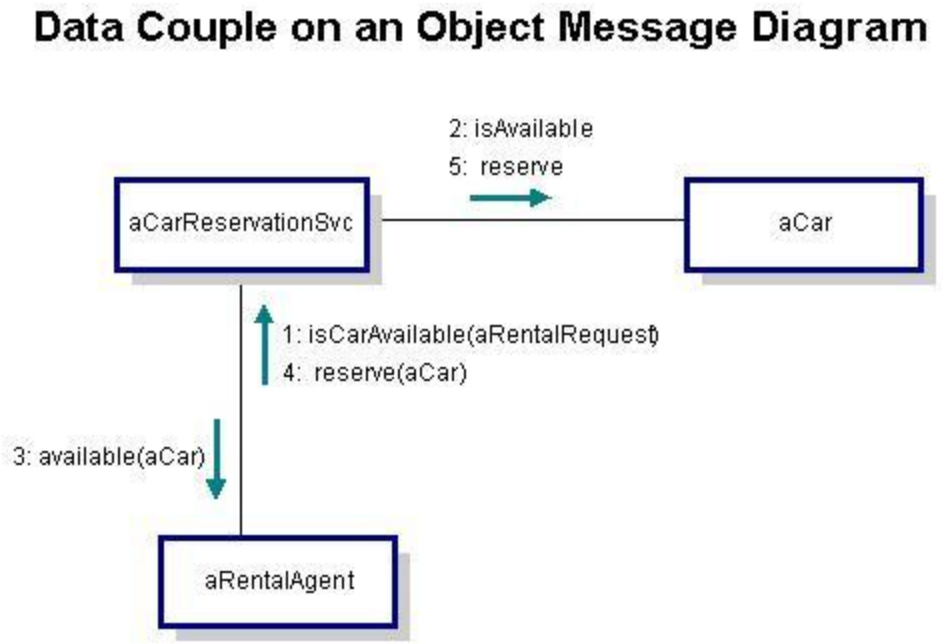
Embed Data Quality in Data Integration

- Clinical research mandates clean data which is scattered
- SAS Clinical Data Integration brings analysis ready healthcare data
- Brings Healthcare specific clinical integration knowledge to transformation modules
- Reduces development time due to custom coding clinical study
- Modules adheres to standards like CDISC SDTM

Keep Data Structures Simple

- Design simple and generic data structures
- Maintain lossless joins to preserve data dependencies
- Maintain right degree of data coupling with fewer parameterized elements

Keep the coupling simple



(Courtesy Craig Borysowich : Chief Technology Tactician)

Justify Data Quality Investment

- Calculate cost to clean before Data Quality
- Document reduction of errors
- Show decrease of customer service calls
- Sell Data Quality to Top Bosses

Ultimately Treat Data as an Asset

How Do We Manage Data as a Strategic Asset? Invest Dollars

Time to Market, Quality, Customer Satisfaction, Compliance, Regulations, etc,etc, …

And use it as an Investment

- Sell high quality data to customers
- Use it as a strategic advantage

Enterprise Data Quality Drivers

Environment Stewardship is part our DNA

At REI outdoor recreation is our passion. We're equally passionate about protecting and maintaining the places where we hike, climb, cycle, camp, paddle and ski. Advancing environmental stewardship through corporate giving, volunteerism and outreach programs is a key REI imperative in protecting the outdoors for future generations.

We want to be around for 100 more years

- Transforming the Customer Experience (TCE)
- Revolutionizing Customer Communications (RCC)
- Right Product, Right Place, Right Time (R3)
- Transforming Business Solutions (TBS)

Improving our Data Quality

- Retail
  - Provide better understanding of Labor data
  - Support multi-channel strategy
- Logistics
  - Increase knowledge of business advocacy
  - Report card distribution analytics
- Online
  - Understand voice of the customer
  - Customer Insights
- Marketing
  - Retention information
  - Customer Analytics
- Finance
  - Understand key sales data for Sales Planning, Traffic Vs Order
  - Taxonomy to improve Sales Planning process

Improving our Data Quality

- Gear and Apparel
  - Track material by country of origin, material type
  - Analytics for trend spotting
- Direct Sales
  - Multi-year view of customer data
  - Total customer membership data
- Merchandise
  - Track customers by medium, warm and affluent
  - Customer data for assortment planning of stores
- Adventure Travel
  - Track customer data for adventure travel up sell
  - Improve data quality of adventure travel data

References

1. Dependencies revisited for improving data quality – Wenfei Fan (Symposium on Principles of Database System) ACM SIGMOD-SIAGACT-SIGART

2. Quality Control Handbook – Joseph Moses Juran, New York, McGraw-Hill, 1951

3. Data Quality (The Kluwer International Series on Advances in Database Systems Volume 23) – Richard Wang, Yang Lee, Mostapha Ziad

4. David Loshin – White Paper – Data Warehouse ROI (Knowledge Integrity/Informatica)

5. A Framework for Analysis of Data Quality Research - Richard Y. Wang, Veda C. Storey, and Christopher P. Firth

6. http://www.sas.com/resources/factsheet/sas-clinical-data-integration-factsheet.pdf

Questions ?

jay_barua@hotmail.com
503-804-1633

Thank You !