You’re the new CDO, now what?

Joe Caserta
President
Caserta Concepts
Caserta Timeline

- **1986**: Began consulting database programming and data modeling
- **1996**: Data Analysis, Data Warehousing and Business Intelligence since 1996
- **2001**: Founded Caserta Concepts in NYC
- **2004**: Co-author, with Ralph Kimball, *The Data Warehouse ETL Toolkit* (Wiley)
- **2009**: Launched Web log analytics solution published in Intelligent Enterprise
- **2012**: Laser focus on extending Data Analytics with Big Data solutions
- **2013**: Dedicated to Data Governance Techniques on Big Data (Innovation)
- **2014**: Top 20 Most Powerful Big Data consulting firms
- **2016**: Awarded Fastest Growing Big Data Companies 2016

---

Established **best practices** for big data ecosystem implementations

Launched **Data Science**, Data Interaction and Cloud practices

Launched **Big Data** practice

Awarded **Top 20 Big Data Companies 2016**

Established best practices for big data ecosystem implementations

Launched Data Science, Data Interaction and Cloud practices

Launched Big Data practice

Founded Caserta Concepts in NYC

Began consulting database programming and data modeling
About Caserta Concepts

• **Consulting**  Data Innovation and Modern Data Engineering
  • Award-winning company
  • Internationally recognized work force
  • Strategy, Architecture, Implementation, Governance

• **Innovation**  Partner
  • Strategic Consulting
  • Advanced Architecture
  • Build & Deploy

• **Leader**  in Enterprise Data Solutions
  • Big Data Analytics
  • Data Warehousing
  • Business Intelligence
# MITCDOIQ

## Technology Partners

<table>
<thead>
<tr>
<th>IBM</th>
<th>Collibra</th>
<th>Cloudera</th>
<th>Hortonworks</th>
<th>MapR</th>
<th>Amazon Web Services</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alation</td>
<td>Tableau</td>
<td>Talend</td>
<td>TIBCO Jaspersoft</td>
<td>Lucidworks</td>
<td>Snowflake</td>
</tr>
<tr>
<td>Datastax</td>
<td>Neo4j</td>
<td>Q2Action</td>
<td>Syncsort</td>
<td>Datameer</td>
<td>Basho</td>
</tr>
<tr>
<td>Revolution Analytics</td>
<td>Cisco</td>
<td>Dataguisepitney bowes</td>
<td>Databricks</td>
<td>Microsoft Azure</td>
<td></td>
</tr>
</tbody>
</table>
A Day in the Life of Data

Business: “I need to analyze some new data”
- IT collects requirements
- Creates normalized and/or dimensional data models
- Profiles and conforms the data
- Sophisticated ETL programs and quality standards
- Loads it into data models
- Builds a BI semantic layer
- Creates a dashboard and reports

..and then you can access your data 3-6 months later to see if it has value!

• Onboarding new data is difficult!
• Rigid Structures and Data Governance
• Disconnected/removed from business
Houston, we have a Problem: Data Sprawl

• There is one application for every 5-10 employees generating copies of the same files leading to massive amounts of duplicate idle data strewn all across the enterprise.
  - Michael Vizard, ITBusinessEdge.com

• Employees spend 35% of their work time searching for information... finding what they seek 50% of the time or less.
  - “The High Cost of Not Finding Information,” IDC
Dealing with the explosion of data sources and data types is an organizational priority.

The need to simplify information is driving significant change in organizations: 43 percent of all organizations currently are making changes to how they design and deploy information and another 37 percent are planning to make changes.
The Evolution of Enterprise Data

Traditional EDW

Big Data Lake

NoSQL Databases

Spark

Cloud

Python

Hadoop Distributed File System (HDFS)

Horizontally Scalable Environment - Optimized for Analytics

ETL

Traditional BI

Ad-Hoc/Canned Reporting

Big Data Analytics

Data Science

Ad-Hoc Query

Canned Reporting

Sales

Marketing

Finance

Others...

EMR/EHR

http://www.

#MITCDOIQ

@joe_Caserta
What’s Old is New Again

• Before Data Warehouse Governance
  • Users trying to produce reports from raw source data
  • No Data Conformance
  • No Master Data Management
  • No Data Quality processes
  • No Trust: Two analysts were almost guaranteed to come up with two different sets of numbers!

• Before Big Data Lake Governance
  • We can put “anything” in Hadoop
  • We can analyze anything
  • We’re scientists, we don’t need IT, we make the rules

• Rule #1: Dumping data into Hadoop with no repeatable process, procedure, or data governance will create a mess

• Rule #2: Information harvested from an ungoverned systems will take us back to the old days: No Trust = Not Actionable
## Data Governance

<table>
<thead>
<tr>
<th>Block</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Organization</strong></td>
<td>This is the ‘people’ part. Establishing Enterprise Data Council, Data Stewards, etc.</td>
</tr>
<tr>
<td><strong>Metadata</strong></td>
<td>Definitions, lineage (where does this data come from), business definitions, technical metadata</td>
</tr>
<tr>
<td><strong>Privacy/Security</strong></td>
<td>Identify and control sensitive data, regulatory compliance</td>
</tr>
<tr>
<td><strong>Data Quality and Monitoring</strong></td>
<td>Data must be complete and correct. Measure, improve, certify</td>
</tr>
<tr>
<td><strong>Business Process Integration</strong></td>
<td>Policies around data frequency, source availability, etc.</td>
</tr>
<tr>
<td><strong>Master Data Management</strong></td>
<td>Ensure consistent business critical data i.e. Members, Providers, Agents, etc.</td>
</tr>
<tr>
<td><strong>Information Lifecycle Management (ILM)</strong></td>
<td>Data retention, purge schedule, storage/archiving</td>
</tr>
</tbody>
</table>
| Organization                                                                 | • Add Big Data to overall framework and assign responsibility  
|                                                                             | • Add data scientists to the Stewardship program  
|                                                                             | • Assign stewards to new data sets (twitter, call center logs, etc.) |
| Metadata                                                                    | • Larger scale  
|                                                                             | • New datatypes  
|                                                                             | • Integrate with Hive Metastore, HCatalog, home grown tables |
| Privacy/Security                                                             | • Data detection and masking on unstructured data upon ingest |
| Data Quality and Monitoring                                                  | • Data Quality and Monitoring (probably home grown, drools?)  
|                                                                             | • Quality checks not only SQL: machine learning, Pig and Map Reduce  
|                                                                             | • Acting on large dataset quality checks may require distribution |
| Business Process Integration                                                 | • Near-zero latency, DevOps, Core component of business operations |
| Master Data Management                                                       | • Graph databases are more flexible than relational  
|                                                                             | • Lower latency service required  
|                                                                             | • Distributed data quality and matching algorithms |
| Information Lifecycle Management (ILM)                                       | • Secure and mask multiple data types (not just tabular)  
|                                                                             | • Deletes are more uncommon (unless there is regulatory requirement)  
|                                                                             | • Take advantage of compression and archiving (like AWS Glacier) |
Why we need a Data Organization

Forces for Change
- Global economics
- Intensity of competition
- Reduce costs
- Move to cross-functional teams
- New executive leadership
- Speed of technical change
- Social trends and changes

Forces Resisting Change
- Period of time in present role
- Status & perks of office/dept under threat
- No apparent reasons for proposed changes
- Lack of understanding of proposed changes
- Fear of inability to cope with new technology
- Concern over job security

Status Quo
What is required of the CDO?

• Provide a single point of accountability for data initiatives and issues
• Innovate ways to use existing data
• Enrich and augment data by combining internal and external sources
• Support efficient and agile analytics through training and templates
• Evangelize a data vision for the organization
• Support & enforce data governance policies via outreach, training & tools
• Monitor and enforce data quality in collaboration with data owners
• Monitor and enforce data security along with Legal/Security/Compliance
• Work with IT to develop/maintain an enterprise repository of strategic data
• Set standards for analytical reporting and generate data insights
The Whole Brain Challenge

Analytics Oriented CDO
- Data Science
- Research

Operations Oriented CDO
- Shared Services,
- Board-level Executive

Revenue Oriented CDO
- Revenue Goals
- Monetizing Data

Process Oriented CDO
- Risk
- Compliance
# Data Organization Roles

<table>
<thead>
<tr>
<th>Role</th>
<th>Responsibilities</th>
</tr>
</thead>
</table>
| **VP, Data**                | • Create and evangelize vision, strategy, and mission statement  
• Create, communicate, and enforce policies, procedures, and processes  
• Plan, prioritize, and project manage data initiatives  
• Prepare & maintain budget for staff, infrastructure, services, tools & training  
• Innovate ways to use existing data  
• Enrich and augment data by combining internal and external sources  
• Protection – ensuring data privacy and security  
• Upkeep – managing the health of data under governance |
| **Data Governance Lead**    | • Represent business interests across departments  
• Prioritize and manage data requests and remediation efforts  
• Identify pockets of business, technical, and data expertise  
• Socialize policies and support programs |
| **Data Stewards**           | • Receive, manage, prioritize and track data quality issues  
• Proactively lead data quality monitoring of high value data  
• Identify, train, and manage critical data sources  
• Ensure remediation efforts follow change management policies  
• Assist in management and maintenance of master data |
| **Data Librarian**          | • Track and manage data related assets (sources, metadata, business glossary, data lineage)  
• Track and manage common queries with embedded business logic  
• Track and manage canned reports (to prevent duplication)  
• Track and manage standard reports and dashboard templates  
• Track internal and external data and tool experts  
• Socialize content being managed with user community  
• Manage the Data Governance knowledge repository |
Additional CDO Responsibilities

Modern Data Engineering/Data Preparation
- Programming/Storage
- Data Quality
- Visualization

Advanced Mathematics/Statistics
- Algorithms
- A/B Testing

Domain Knowledge/Business Expertise
- Data and Outcome Sensibility

Data Science and Insights
Corporate Data Pyramid

Usage Pattern

- Arbitrary/Ad-hoc Queries and Reporting
- Munging, Blending Machine Learning
- Organize, Define, Complete
- Data Ingestion

Data Governance

- Fully Governed (trusted)
- Data Quality and Monitoring
  - Metadata, ILM, Security
- Data Catalog
- Data Integration
- Metadata, ILM, Security

Landing Area – Source Data in “Full Fidelity”

Data Lake – Integrated Sandbox

Data Science Workspace

Big Data Warehouse
Caution: Assembly Required

- Some of the most hopeful tools are brand new or in incubation
- Enterprise big data implementations typically combine products with custom built components

People, Processes and Business commitment are still critical!
CDO Success in Summary

Data Centric, Technology Enabled, Business Focused

**Streamline Processes**
- Self-service, reduce ongoing dependency on IT
- Automate Workflows

**Business Definitions**
- Identification of KPI's
- Iterative Process – definitions mature over time
- Tools provide user-centric experience

**Automation**
- Data Discovery
- Data Profiling
- Workflows
- Data Quality
- Automated ILM

**Roles**
- CDO
- Data Governance Council
- Data Stewardship Team
- Business SME’s
- Data Scientists for Insights

**Architecture**
- Consolidated view of data
- Flexibility for future growth
- Viewable Everywhere

**Metrics**
- Gauge overall governance of data
- Data Quality reporting
- Issue Tracking

#MITCDOIQ
@joe_Caserta
Questions & (hopefully) Answers

Joe Caserta
President, Caserta Concepts
joe@casertaconcepts.com
@joe_Caserta

Thank you
Sample Data Architecture