

5th Annual

Cloudera Government Forum



cloudera

Perspectives on Ethical Big Data Governance

Murthy Mathiprakasam, Principal Product Marketing Manager, Informatica
Steven Totman, Financial Services Industry Lead, Cloudera

Cloudera
Government
Forum

Governance of Big Data?

cloudera



What is Metadata

So what is “Metadata”?



Metadata enables you to put context and meaning to things.
It is generated and consumed by **every** organization and software product.

So what is “Metadata”?



Metadata enables you to put context and meaning to things.

It is generated and consumed by **every** organization and software product.

Enterprise Metadata

Enables Data Governance

B

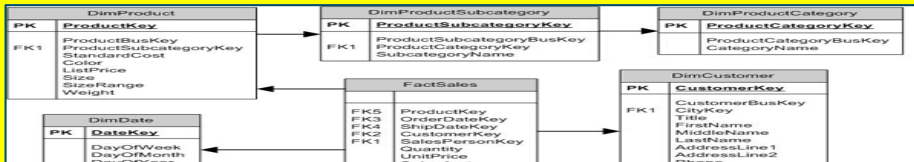
Business Metadata



Business Glossary
Enterprise Taxonomy
Ontology

T

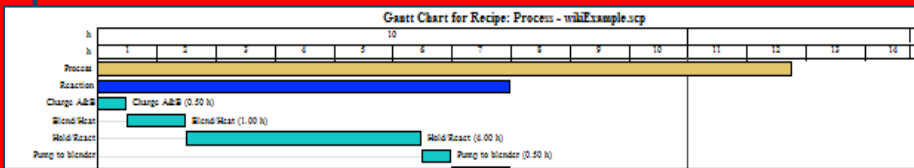
Technical Metadata



Database Schema
File Definition
Data Flow Design
BI Report Definition
Data Model

O

Operational Metadata



Job Run-time Stats
Report run information
Hardware Usage
Scheduler Stats

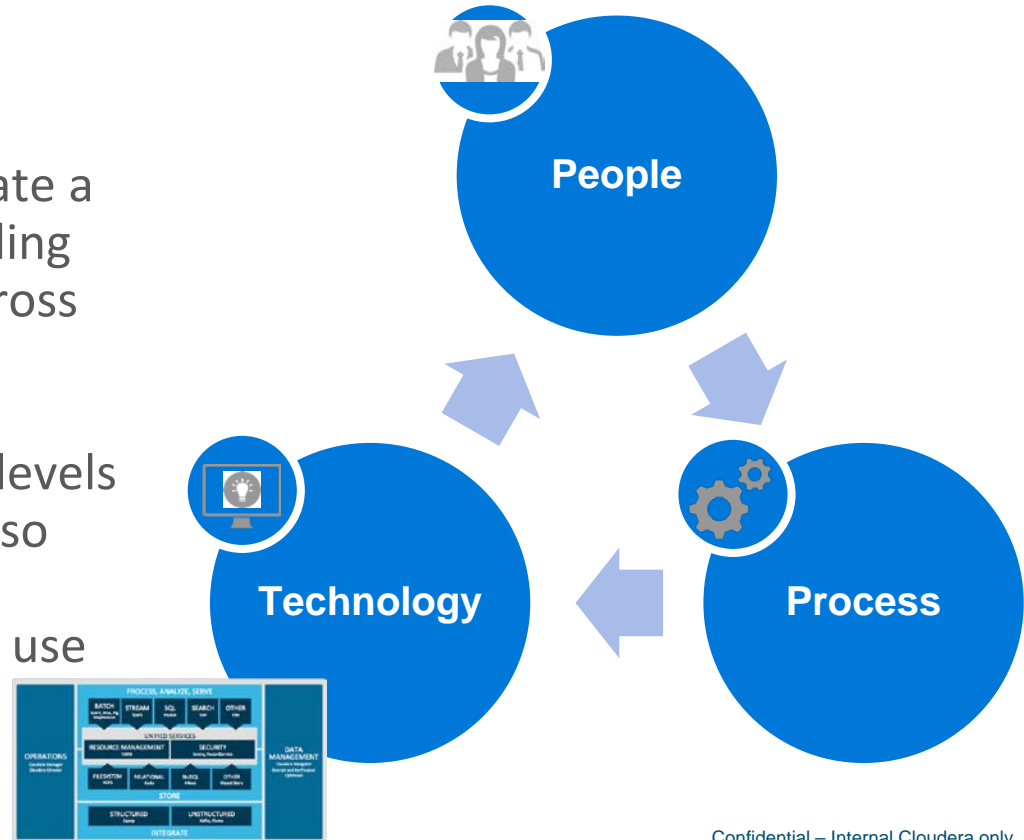
Data Lineage
Impact Analysis
Topology Understanding
SOX Compliance
Auditing

Literally, “data about data” that describes your company’s information from both a business and a technical perspective

What is Data Governance

What is Data Governance?

- Encompasses the **People**, **Processes** and **Information Technology** required to create a consistent and proper handling of an organization's data across the business enterprise
- Goals may be defined at all levels of the enterprise and doing so may aid in acceptance of processes by those who will use them



Data Ethics

Jake Sorofman - Gartner: Don't be creepy

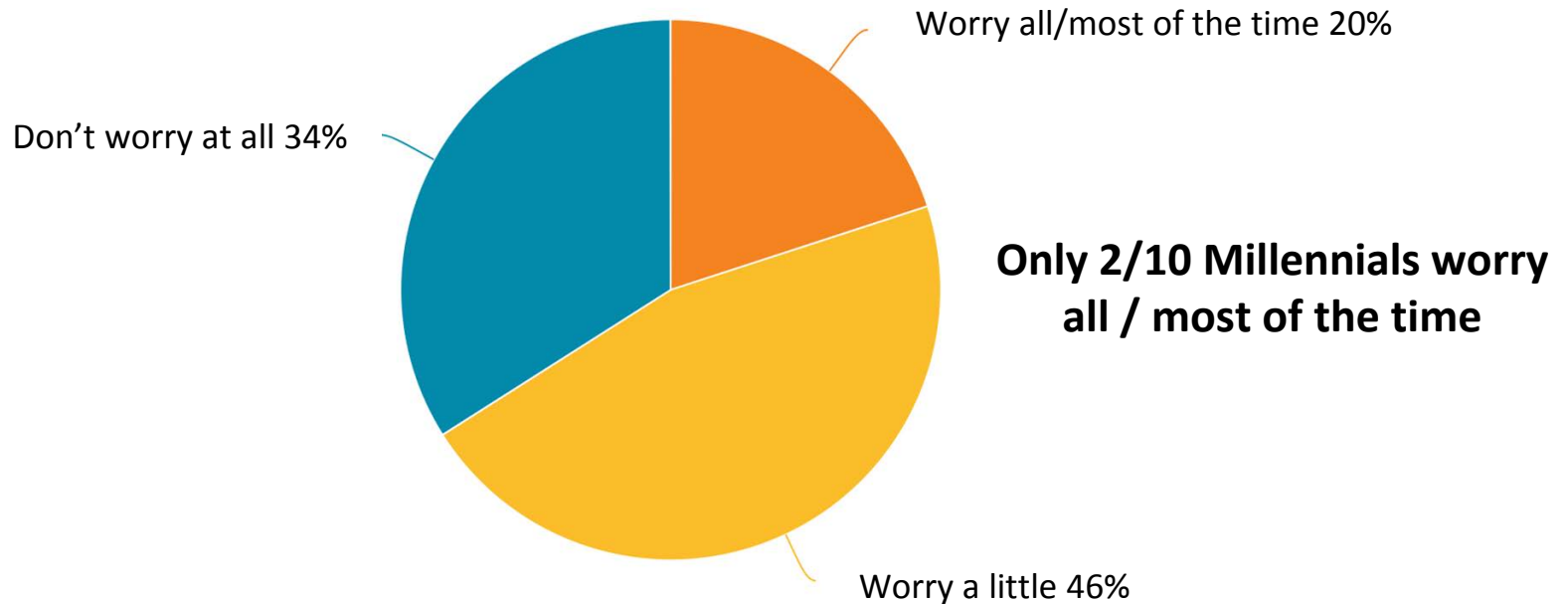
I think personalization has the potential for unmistakable good—for both consumers and brands. But the question remains: *how do you hew the line between personalized and downright creepy?*

Comic book prophet Stan Lee told us **“With great power comes great responsibility.”**



Millennials are not very worried about their privacy online - Digital lives of millennials survey

Question: how much do you worry, if at all, about information about you being available online?

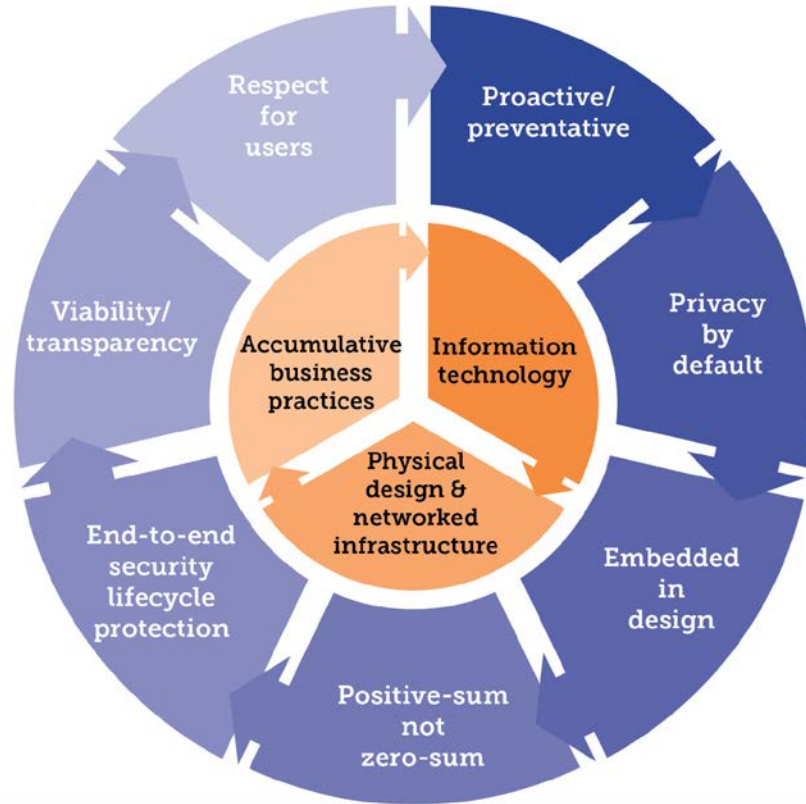


Ethical Data Usage



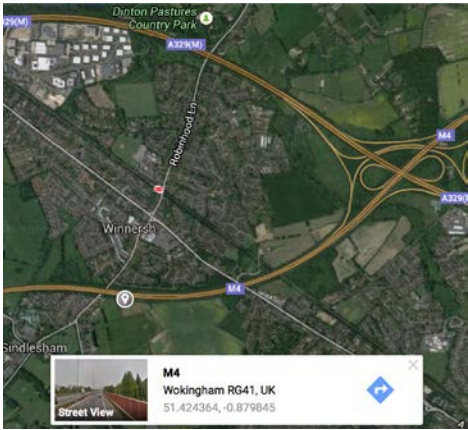
Good way to evaluate -
What would mum and dad
think?

Privacy by Design



Based on 7 Foundational Principles, *PbD* was first developed by Ontario's Information and Privacy Commissioner, [Dr. Ann Cavoukian](#), in the 1990s.

Real Life Case Study: Network Intelligence



These common policy guidelines from the “toolkit” ought to inform data protection and usage

1. Information should be collected in a legal manner for a specific legitimate purpose.
2. A risk assessment of the data collection activity should inform the design of the data collection process.
3. Data subjects should provide informed consent for the data collection process.
4. Data should be secured to prevent unintended uses.
5. Data should not be held indefinitely, and should be destroyed when no longer needed.
6. Affected people should be able to request information about what personal data an organization holds about them.
7. Duplication of information collection efforts should be avoided.



It's Becoming Harder to Protect Your Organization



Threat Surface Expanding

16 billion connected devices generating more data



Attacks are Increasing

Attacker sophistication has increased leading to 250% more successful attacks

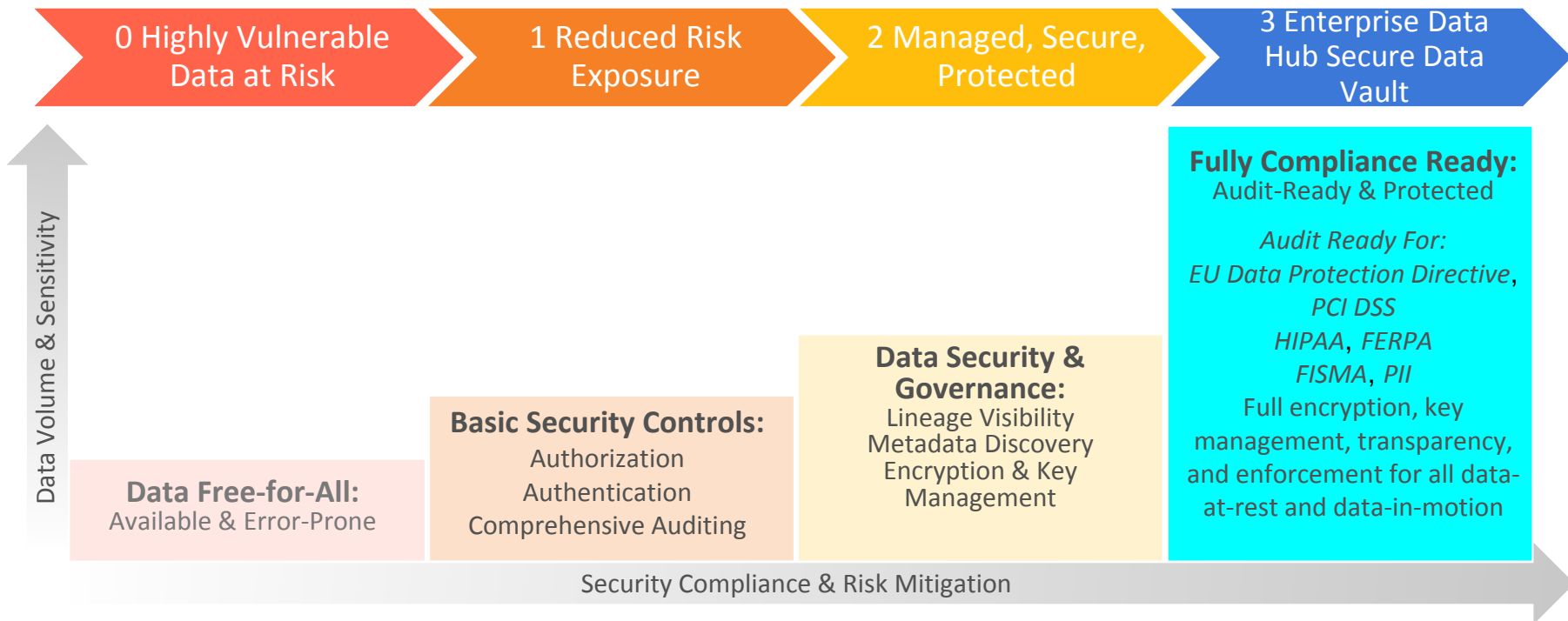


Threats Are Highly Adaptive

Protection against attacks with known signatures no longer sufficient

Start with the Hadoop Security Maturity Model

Achieve Scale and Cost Effectiveness via a Secure Data Vault

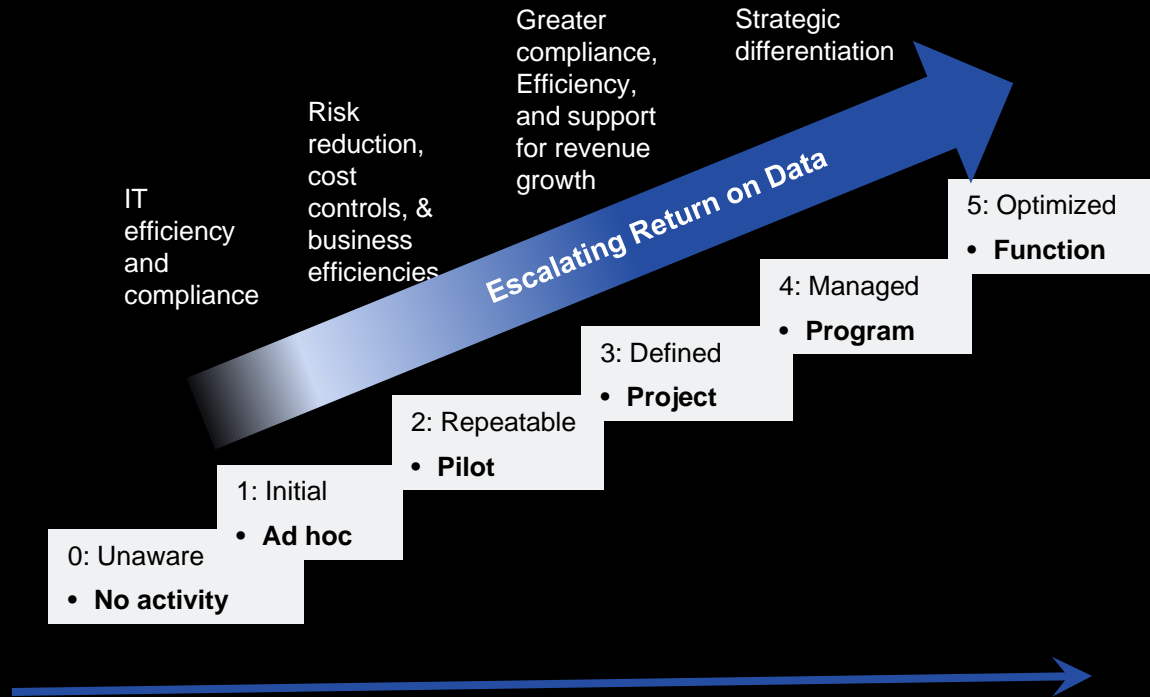


A large elephant is the central focus, its trunk pointing downwards. The background is a dark, abstract space filled with glowing, multi-colored lines (yellow, red, blue) that suggest data flow. Scattered throughout this background are various city names and numbers, such as 'San Antonio', 'Houston', 'Fort Myers', 'Tampa', 'Long Beach', 'Memphis', 'Germans', 'Hushuig', 'Wills Point', '3357,880', '80,500,606', '\$43', '\$375,900', '95,820', '123,628', '777', '27,900', and '26'.

Perspectives on Ethical Big Data Governance

Murthy Mathiprakasham
mmathipra@informatica.com
@mmathiprakasham

Big Data Can Be An Asset or a Liability



Data Governance Maturity

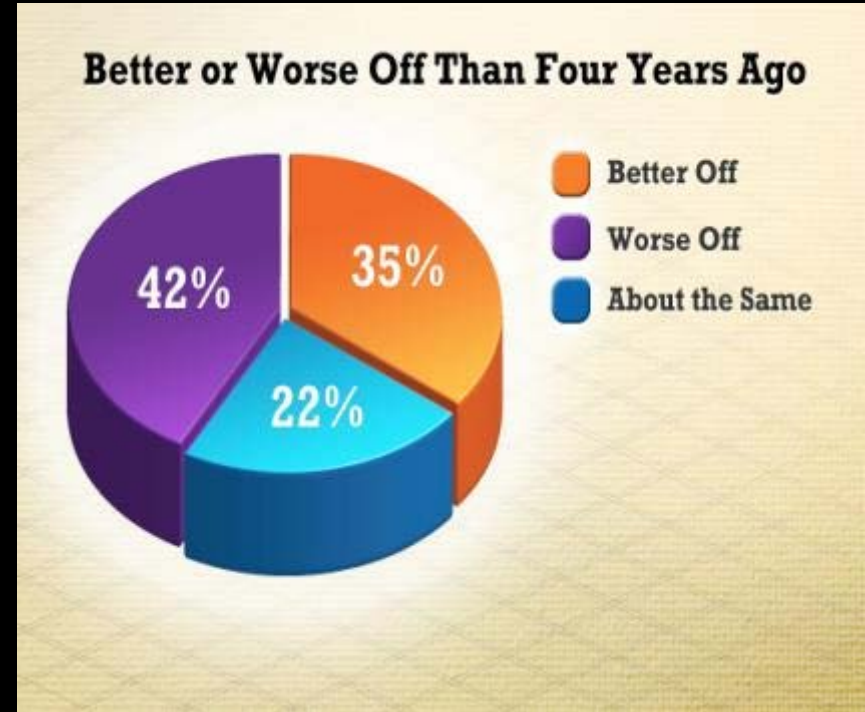
Data is an enterprise resource like any other resource (financial capital, human capital, etc)

- **Passively**, enterprises must comply with internal controls and external regulations
- **Strategically**, enterprises can drive greater analytical value with governed assets

Is Data Becoming Too Pervasive?



A screenshot of a mobile application interface showing a sponsored post. The status bar at the top shows AT&T, 9:46 PM, and 65% battery. The post is titled "Suggested Post" and features a logo for "I Am Shapiro" with a "Sponsored" tag. The text of the post reads: "Is your last name Shapiro? If so, this shirt is for you! Limited Edition Team Shapiro Lifetime Member shirt. Available as a t-shirt, women's relaxed fit, v-neck, long sleeve, or hoodie." Below the text is a button that says "Select your style and or... Continue Reading". The main image shows a black t-shirt with "TEAM SHAPIRO" printed on it. Below the image, it says "Click To Buy for Just \$19.99 (Save \$10 Today)" and "limited time offer don't miss out". At the bottom of the post, it shows "18 Likes 7 Comments". The navigation bar at the bottom includes icons for News Feed, Requests, Messages, Notifications, and More.



Your Mission

Define and Enforce an Ethical
Policy of Governance That Delivers
A Great Public Outcomes While
Ensuring Trust By Understanding,
Protecting, and Tracking Your Data

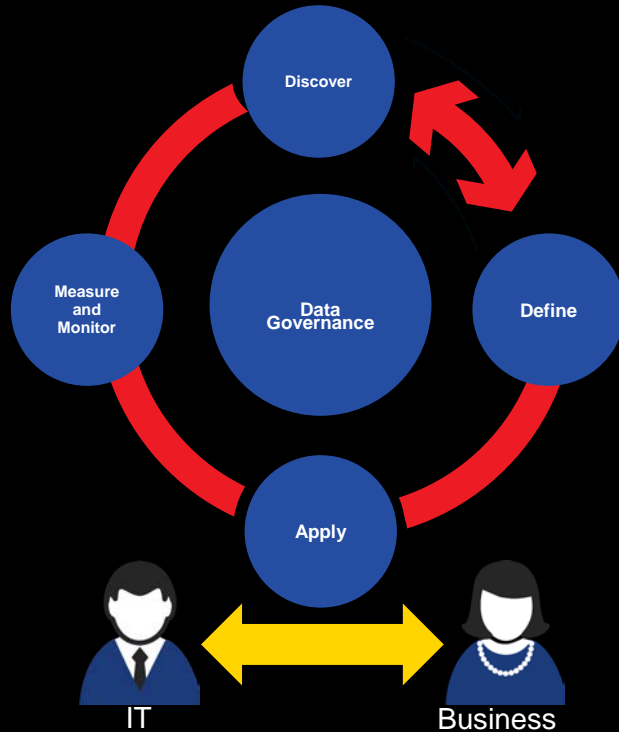
Keys To Success



Train Your People on a Defined Code of Ethics

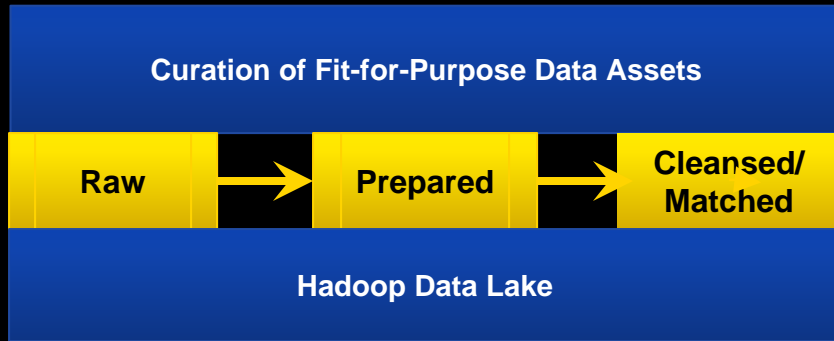


Best Practices for Big Data Governance - 1



- Define Big Data Governance Process – Roles, Standards, Taxonomies, Business Glossaries
- Discover Big Data Assets
- Apply, Measure, and Monitor Enforcement
- Leverage IT & Business Collaboration To Balance Big Data Governance Objectives With Agility

Best Practices for Big Data Governance - 2



- Quality and Governance are not fixed points
- Use flexibility of Hadoop alongside curation process and technology for fit-for-purpose data assets to get right data to right people at the right time

What Technology Challenges Are Customers Facing?

“too many data silos making it impossible to know what data can be trusted”

Pete, Chief Data Officer

“need to ensure confidence in data integrity, accuracy, and timeliness”

Ron, VP Global Information Systems

“regulations have become very strict and very precise – lots of gaps in the quality of the data”

Christine, Manager Data Management

“need code re-usability and code maintainability”

Ben, Director of Platform Architecture

“transforming data management from a labor intensive, qualitative approach to a systematic approach...to classify data and understand lineage”

Ned, Senior Vice President

Big Data Cannot Be Tackled Manually



The Race to Business Value Will Not Be Won By Hand



More
Volume



More
Variety



More
Velocity



More Data
Platforms



More Data
Consumers



More Data
Silos

Big Data Is Difficult To Trust



Changing

Needs for Quality

Same data used for
multiple purposes



Hidden

Relationships

Everything and everyone
is interconnected



Magnified

Trust Issues

New sources of
external data

And Regulations And Controls Are Harder To Meet

SOX

PCI

HIPAA

FISMA

ISO

GLBA

NIST



Perimeter Security Is Insufficient



Perimeter security: Outside in security

- Not if, but when
- Network focused
- Attacks will only grow



Big Data: Bigger Risk



Sensitive
Data




Security
Exposure




- An exponential attack surface
- With exponential risks

In the Public Sector, This Means People Can Suffer Impact

 Can't make **comprehensive decisions** based on all of the available data

 Can't make **accurate decisions** based on high quality and secure data

 Can't make **timely decisions** based on fresh and up-to-date data

 Can't **operationalize data delivery** to fuel decisions repeatably and scalably

Informatica Big Data Management 10.1



“Project Sonoma”



**Find & Access
Any Data Centrally**



**Discover Data
Relationships**



**Prepare & Share
Relevant Data**



**Operationalize Business
Insights Quickly**

PILLAR 1 Big Data Integration

- Optimized Execution & Flexible Deployment
- 100's of Pre-built Transforms, Connectors & Parsers
- Flexible Deployment to Cloud

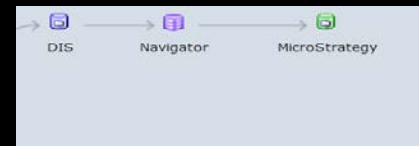
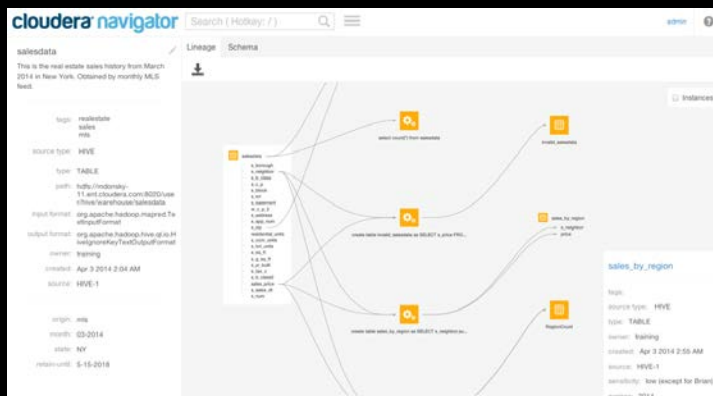
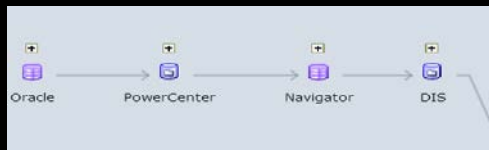
PILLAR 2 Big Data Governance & Quality

- Business Glossary
- Profiling and Data Quality
- 360° Relationship Views

PILLAR 3 Big Data Security

- Discover sensitive data in Hive
- Dynamic data masking for Hive
- Proliferation Analysis

Big Data Management Meets Data Governance



Informatica Big Data Management + Cloudera Navigator

- Seamless integration with Cloudera Enterprise
- Track the end-to-end lineage of data before, during, and after Hadoop processing with metadata management and lifecycle management
- Use data profiling capabilities and pre-built business rules to evaluate and mitigate data quality
- Comprehensive platform for managing technical, business, and operational metadata

Your Mission

Define and Enforce an Ethical
Policy of Governance That Delivers
A Great Public Outcomes While
Ensuring Trust By Understanding,
Protecting, and Tracking Your Data



Great Transportation

- **Aspiration:** Florida Turnpike sought to improve emergency preparedness and improve the prepaid toll program
- **Challenge:** Data collection took over one month leading to faulty analytics
- **Outcome:** “Timely and accurate traffic, revenue, and participation reports help management make good choices that will eventually result in saving money.”
- — Bob Hartmann, IT Director, Florida Turnpike Enterprise



Great Environment

- **Aspiration:** US Geological Service sought to improve the quality of water in the United States
- **Challenge:** Collect distributed data and build a centralized water quality dataset
- **Outcome:** “We chose Informatica as our data integration solution because of its maturity, wide range of features, ease of use and industrial strength, integrated architecture.”
- — Harry House, Data Warehouse Practice Leader, USGS



Great Education

- **Aspiration:** Rochester Institute of Technology sought to understand how it could improve student enrollment, student housing, and student retention
- **Challenge:** Data was in disparate systems
- **Outcome:** “We’re becoming myth busters. Informatica provides timely, accurate information we need to spot trends, improve the quality of our academic learning, and reduce attrition.”
- — Kim Sowers, Director of Application Development, Rochester Institute of Technology



Great Healthcare

- **Aspiration:** Utah Dept of Health sought to process healthcare claims faster and improve public health
- **Challenge:** Manual effort to track and link claims data over time
- **Outcome:** “We see the Informatica as absolutely essential to everything that we want to do, not only to meet our mandate for the All Payer Database,”
- — Dr. Keely Cofrin Allen, Director, Office of Health Care Statistics, State of Utah Department of Health

Getting Started is Easy

1.

Inventory Data &
Understand Related
Legal frameworks



2.

Define and publish
usage guidelines &
privacy policies



3.

Contact us or a Partner
to Start a POC



Define & share “what is legal and what is right” - for your organization

Final Thought

Ethical Big Data Governance today is
“**Like kissing in the school yard**”

- Everybody seems to be talking about it
- Very few people are actually doing it
- Even fewer are doing it well

