



Advancing all things data.

# Big Data Management Best Practices for Data Lakes

**Philip Russom, Ph.D.**

Senior Research Director, TDWI

October 27, 2016

# Sponsor



**informatica**

Put potential to work.™

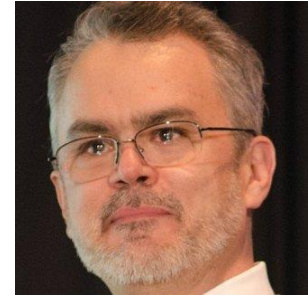
# Speakers



**Philip Russom**  
Senior Research Director  
for Data Management,  
TDWI



**Murthy Mathiprakasam**  
Director of Product Marketing,  
Big Data Products,  
Informatica



**Tavo De Leon**  
Global Leader,  
Big Data Strategy & Solutions,  
Cognizant

# Agenda

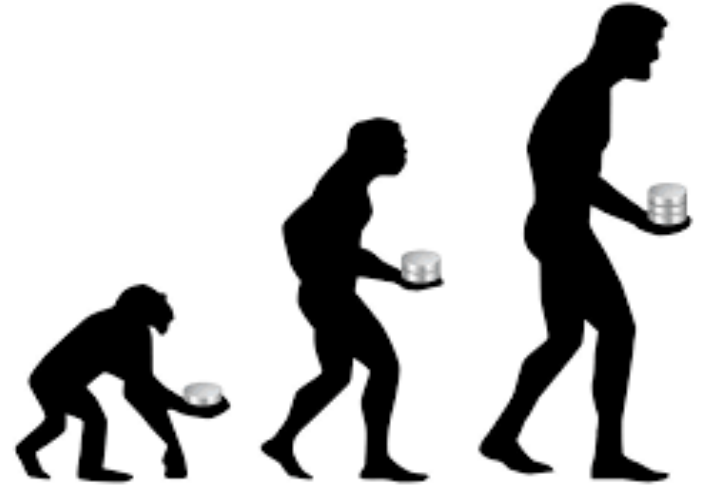


- Background for Data Lakes
  - Evolving data, management, biz use
  - Defining Data Lakes & their benefits
  - Hadoop's role in Data Lakes
- Seven Points
  - Data Management (DM) Best Practices (BPs) for Successful Data Lakes
- Summary
  - I'll draw a "big picture" graphic to summarize and pull together my points

**PLEASE TWEET**  
**@pRussom, @InformaticaCorp,**  
**#TDWI, #DataLakes**

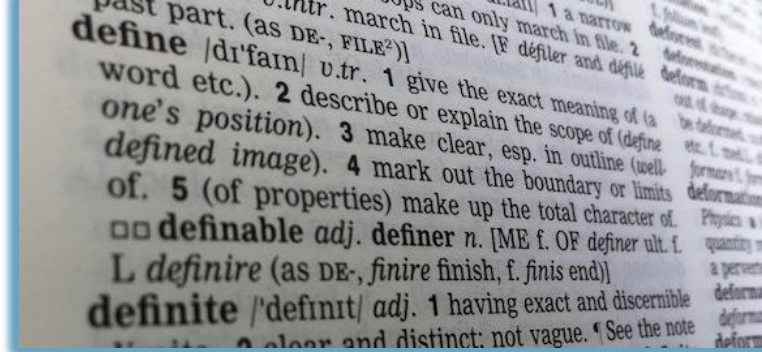
# TRENDS DRIVING Great Changes in Data Management

- Data is evolving
- Data management is evolving
- Business use & leverage of data is evolving

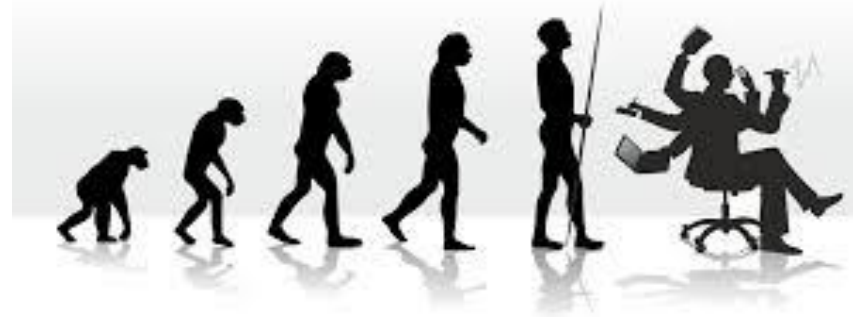


# DEFINING Data Lake

- Handles large volumes of diverse data
  - *For broad data exploration and analytics*
- Ingests data quickly
  - *So data is ready ASAP for exploration, analytics, reporting...*
- Persists data in its original raw detailed state
  - *So analysts have ample raw material they can repurpose later for new analytic questions, business requirements, and other use cases*
- Flexibly supports data management (DM) best practices (BPs) during...
  - *Exploration time, for ad hoc discovery and analytics*
  - *Intermediate stages to prep data for reporting and performance management*
  - *Later in time, as new analytic applications are envisioned*
- Supports multiple use cases in multiple data architectures
  - *Data warehousing, data integration, analytics, reporting; marketing, sales...*



# Okay, but what do I get from a data lake?



- Decisions based on more and better facts
- More complete views of customers and other key business entities
- Turn massive datasets and analytics into competitive advantage
- Business value from non-relational data
- Leveraging big data and new sources, instead of hoarding it
- More “years” of data to analyze; cross-source correlation; exploration...

# Data Lakes and Data Warehouses are Complementary, and They Integrate Well

## Data Warehouse

- RDBMS for rich functionality
- Typically schema on write
- Enterprise-standard data
- Mostly refined data
- Known entities, tracked over time
- Data transformed a priori

## Data Lake

- Hadoop for scalability at low cost
- Typically schema on read
- Fidelity to original format
- Mostly detailed source data
- Raw material to explore & discover
- Repurpose later, as needs arise



# Consider Hadoop

## As a Platform for your Data Lake



- Hadoop isn't required. Lakes can be deployed on:
  - Other file systems or a relational database management system (RDBMS)
- However, Hadoop is a good fit for data lakes:
  - Captures and manages big data at scale
  - Manages multi- and unstructured data
  - Data landing and staging on steroids
  - Repository for detailed source data
  - Processing for analytics, data integration, ETL
  - Inexpensive compared to RDBMSs

# Seven Data Management (DM) Best Practices (BPs) for Data Lakes

1. Biz value from both old and new data
2. Enable data lake best practices with multiple forms of metadata
3. Use metadata to navigate the lake
4. Organize fit-for-purpose data assets
5. Govern the data lake so it won't become a swamp
6. Design and unify your lake for business analytics
7. Build a hybrid data supply chain architecture that includes a data lake on Hadoop

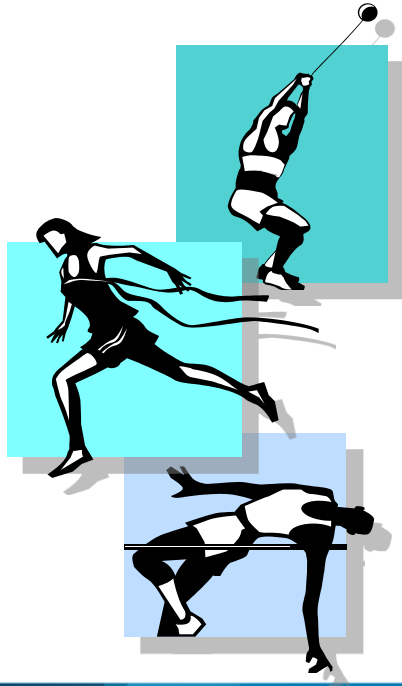


# Biz Value from Both Old & New Data



- Use cases for new big data
  - Sensor data and other machine data
  - Streaming data
  - Human language text
- Use cases involving traditional enterprise data
  - Miscellaneous server logs
  - Selected components of data warehouse architectures
- Use cases for cross-source analytic correlations
  - The marketing data lake; Sales performance lake
  - Healthcare data lake; any analytic data lake

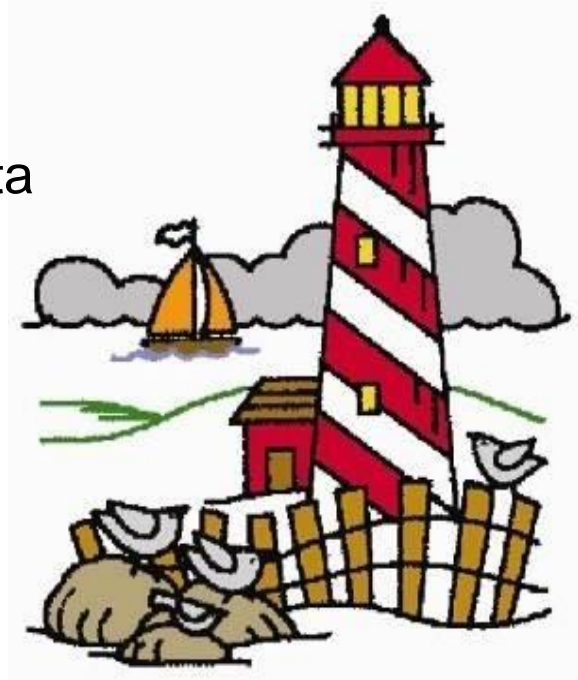
# Enable data lake best practices with Multiple forms of Metadata



- Technical Metadata
  - Nerdy names for automated software access
- Business Metadata
  - Descriptions that humans can understand
  - Gives biz people agile self service w/ lake data
- Operational Metadata
  - Records details about a data access event
  - Enables auditing & security intelligence

# Leverage modern metadata repositories to Navigate the Data Lake

- Enterprise information catalog
  - Complete, unified view of data via metadata
  - Collaboration around data
- Dataset recommendations
  - Automatically suggest datasets that are available, similar, used by other analysts
- Data security intelligence
  - Identifies sensitive data and protects it



# Leverage a Data Lake and DM BPs to Organize Fit-for-Purpose Data Assets



- Dataset gradation into data zones
  - Zones for data landing, exploration, sand boxes, sync w/operational apps, end user consumption...
  - Just enough structure; lake still true to raw source
- Logical structures for the data lake
  - Virtual, federated, views; like a logical DW
- Data integration hub – more than storage
  - Orchestration, pub/sub, architecture for data lake

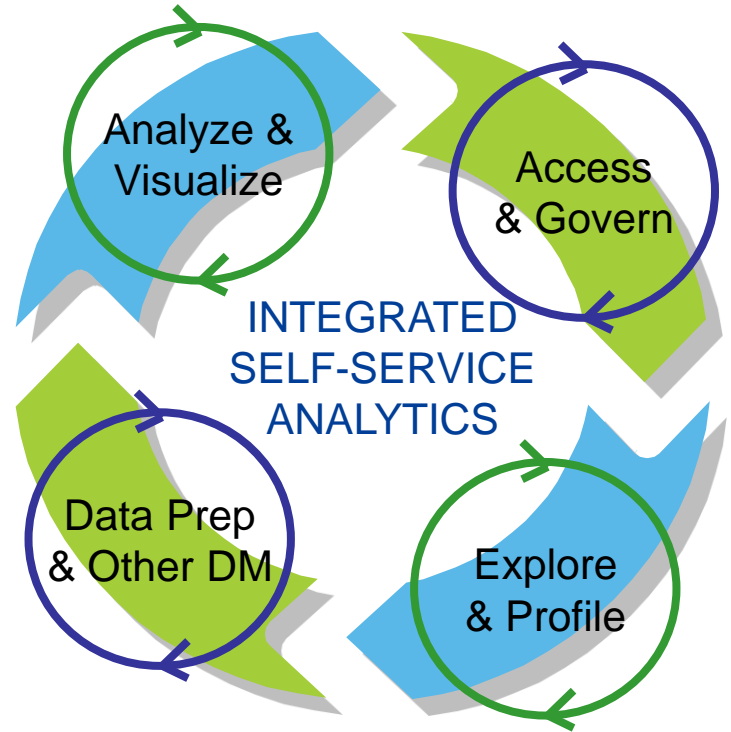
Use policy-based  
data governance so  
**Data Lake**  
won't become  
a Swamp



- Collaborative data governance (DG)
  - Biz, tech & other people collab via DG board
- Data stewardship and curation
  - Data 'owner' guides biz-driven improvements
- Data ingestion
  - Never dump data into a lake; vet & doc it all
- Data lineage
  - Key to users' trust & reducing redundant data
- Data quality
  - Critical for all data, even new data in a lake

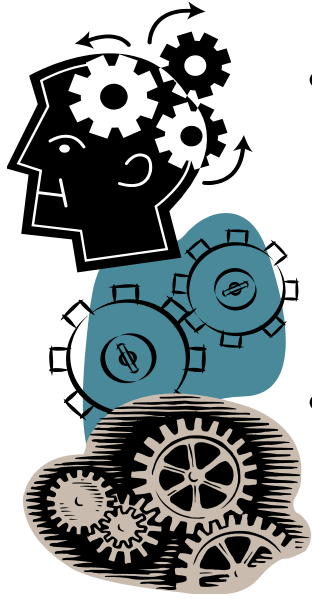
# Enable integrated business analytics with Self-Service Tools for Data Lake Users

- Users need agile self-service analytics
- Involves multiple self-service tasks
  - Data access, exploration, profiling, data prep, visualization, analytics...
- Users want seamless movement from one self-service task to another
  - Provide tools that integrate to enable “integrated self-service analytics” →





# Include a Hadoop-based data lake in your Data Supply Chain

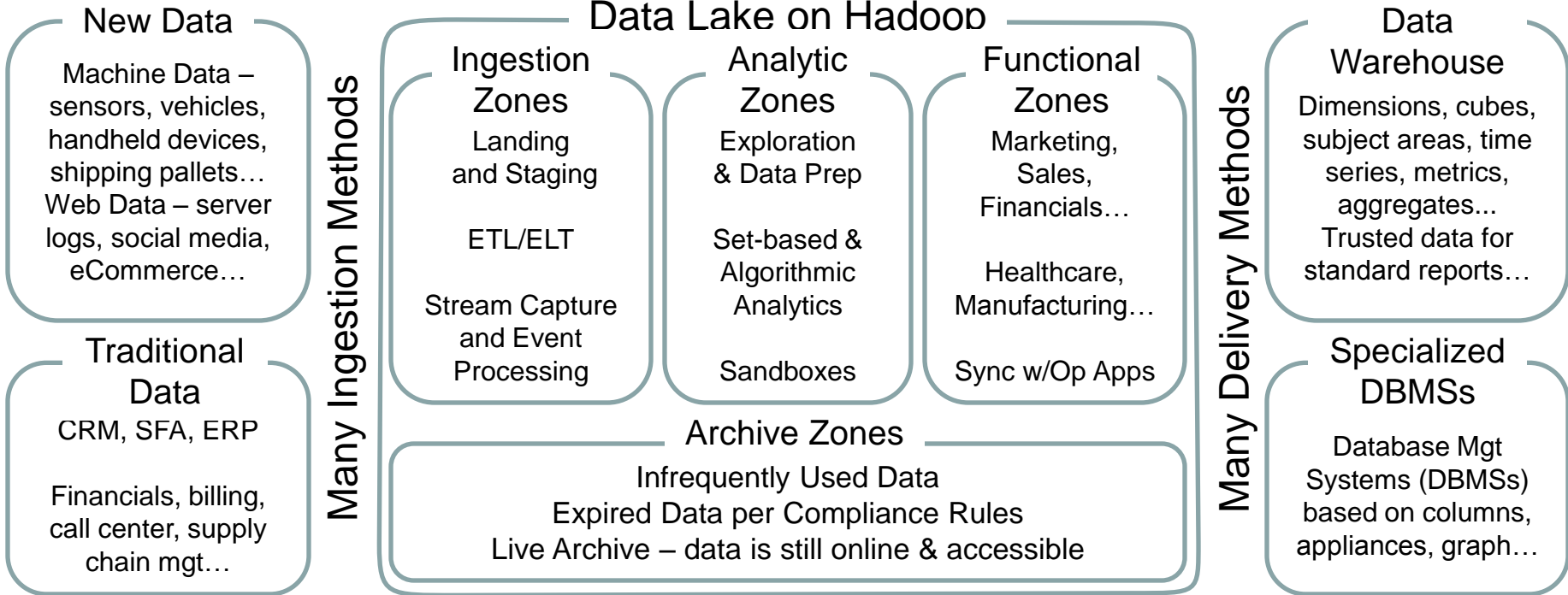


- Data diversification drives platform diversification
- The result is the modern hybrid data ecosystem, which offers challenges:
  - Data travels – a lot – across the multiple data platforms
  - Complex architecture of platforms, datasets, flows, users
- How can you manage a hybrid data ecosystem?
  - Build a **data supply chain** atop a Hadoop-based data lake and related architectural components

# Data Supply Chain for Hybrid Data Ecosystems

Data Integration and Metadata Management Infrastructure

Plus Views: Logical, Virtual, Federated





**informatica**

**Comprehensive  
Data Lake  
Management**



# Who is Informatica?

## World's #1 provider

of data management solutions for Cloud, On-premise, Big Data, and Hybrid environments.

## 7,000+ organizations

around the world turn to Informatica for Data 3.0 solutions to power their business.

**Data Management Leader for 23 Years**



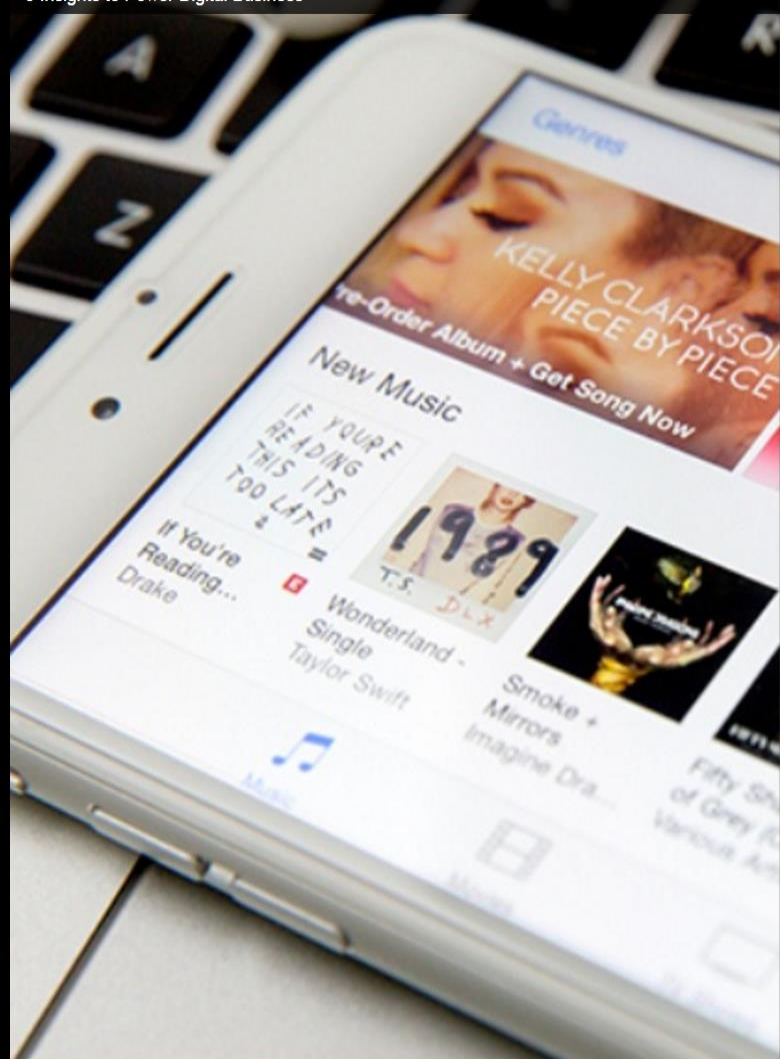
# Who is Cognizant?

## Cognizant

is a leading provider of information technology, consulting, and business process services

**Over 100 WW Development/Delivery Centers**





# Use These 5 Data Sources to Provide a Better Customer Experience

Source: Riding the Seven Waves of Change to the Digital Future, Cognizant

1. Web activity
2. Wearables
3. Smart products
4. Social media
5. Internal legacy systems

# HSBC improved fraud detection by monitoring use of millions of cards!

Source: How Data Fuels Banking's Digital Transformation

Preventing bank fraud before it happens means big savings.



# 65% of consumers don't know where their personal data is stored

Source: Cognizant Center for the Future of Work

Don't play digital hide and seek  
with your customer's data – or  
they may not see you any more.





# Organizations are Facing Challenges Getting Value from Data Lakes

**Insufficient Understanding**

***“Need to bridge information silos and bring data into a central location.”***

*Hamilton, Head of Global Data Strategy*

***“Need access to our data. ..we had data silos. . .this led to analytics silos as well.”***

*Raman, VP Enterprise Analytics Management*

**Unacceptable Wait Times**

***“Prepping and cleaning the data used to take us 2-3 weeks.”***

*Vishal, VP Data Architecture*

***“Need shared data services with a common authoritative set of data assets.”***

*Mike, Chief Data Officer*

**Unknown Trust**

***“Our customer account and contact data was often stale, inaccurate or incomplete.”***

*Barbara, Chief Data Governance Officer*

***“Need to ensure confidence in data integrity, accuracy, and timeliness.”***

*Ron, VP Global Information Systems*

**Inconsistent Delivery**

***“Need code re-usability and code maintainability.”***

*Ben, Director of Platform Architecture*

***“Need to transform from a labor-intensive, qualitative approach to a systematic one...”***

*Ned, SVP Enterprise Data Management*

# Fragmented Approaches Only Create More Complexity

ACQUIRE



Weblogs



Device data



Files



Social



Relational

INGEST

PREPARE

CATALOG

SECURE

GOVERN

ACCESS

CONSUME



Data Mining



Dashboards



Applications



Files

**Hand Coding**  
**One-Off Code Generators**  
**Fragmented Tools**

**NO METADATA INTELLIGENCE**

**LIMITED SUPPORT FOR PROCESSING ENGINES**

**LIMITED SUPPORT FOR INFRASTRUCTURE**

# Informatica's Comprehensive Solution for Data Lakes

ACQUIRE



Weblogs



Device data



Files



Social



Relational

INGEST

Broadest Connectivity

Batch Processing

Stream Processing

PREPARE

Data Parsing

Data Profiling

Data Preparation

CATALOG

Enterprise Data Catalog

Data Lineage

Big Data Relationships

SECURE

Data Security Intelligence

Data Protection

Sensitivity Visualization

GOVERN

Data Mastering

Record Linkage

Business Glossary

ACCESS

Data Quality

Reusable Workflows

Publish / Subscribe

Informatica Data Lake Management

CONSUME



Data Mining



Dashboards



Applications



Files

METADATA INTELLIGENCE

Catalog

Lineage

Search

Recommendations

COMPREHENSIVE SUPPORT FOR DATA PROCESSING

Spark

Spark Streaming

Blaze

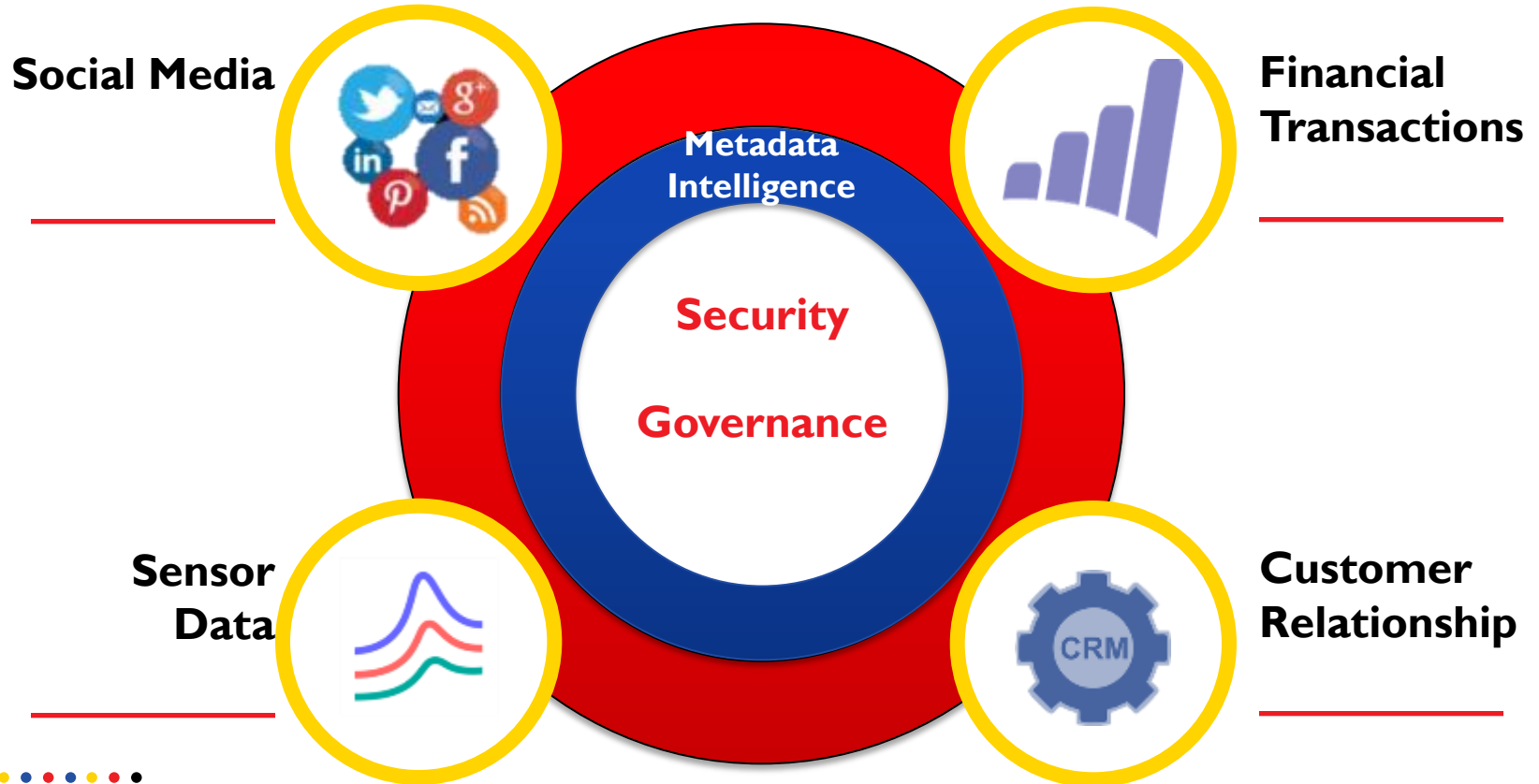
Tez

MapReduce

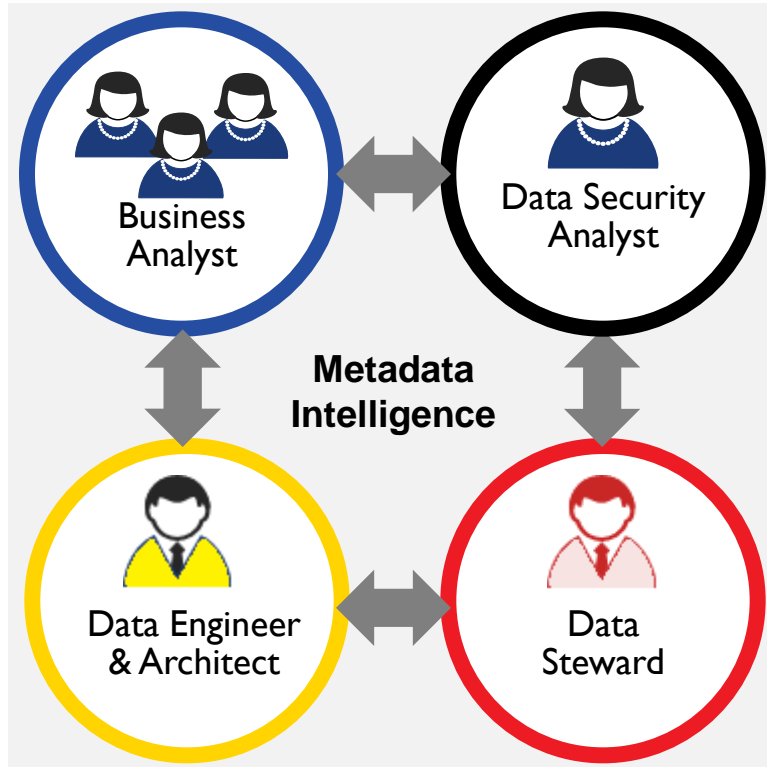
COMPREHENSIVE SUPPORT FOR DATA INFRASTRUCTURE



# Informatica Integrates Data for Analytics, Security, and Governance



# Informatica Integrates Stakeholders



Big Data Management is a “Team Sport”

## Data Steward:

- Provide business glossary for project
- Provide overall context for Analysts

## Business Analyst:


- Define transforms for IT to operationalize
- Innovate new data objects for Stewards

## Data Engineer & Architect:

- Eliminate manual effort for analysts
- Provide technical metadata for project

## Data Security Analyst


- Define security policies and assess risk
- Ensure sensitive data is protected

The background of the slide features a blurred, high-angle view of a modern office interior. In the foreground, several business professionals are silhouetted against a large window. They appear to be in conversation, with some holding briefcases. The window looks out onto a city skyline with several skyscrapers under a clear sky. The lighting is warm, suggesting a sunset or sunrise, with a bright sun flare visible through the window on the right side. The overall atmosphere is professional and dynamic.

“The #1 problem we had was a trust issue. Now we use Informatica as an information management hub to facilitate the movement of trusted data.

We have about 70 applications that contain valuable customer information. We needed to move it into a central location where we could improve the quality of the data to ensure that it was accurate, complete and consistent across our key applications. This was a foundational investment. We needed to gain a 360 customer view across the business.”

-- Informatica Customer

The background of the slide features a silhouette of several business professionals in a modern office setting. They are standing in front of a large glass window that offers a view of a city skyline at sunset or sunrise. The sun is low on the horizon, creating a bright, golden glow that silhouettes the people and the office structure. The reflections of the people and the city are visible on the polished floor.

“This is the first time we’re getting data back to the source system, which is actually really good. And you know, we talk about wanting to do all of this stuff, but it never gets fixed; it’s in perpetuity.

Now, you start helping your business to see their data issues, to make it transparent, to make it visible, and for a company of our size, that’s not easily done.”

-- Informatica Customer

# Thank you to our sponsor



# **informatica**

Put potential to work.™



# Questions?



# Contacting Speakers

- If you have further questions or comments:

Philip Russom, TDWI  
prussom@tdwi.org

Murthy Mathiprakasam, Informatica  
mmathipra@informatica.com