
# Enterprise Architecture Approaches to Big Data

## *Key Concepts and Best EA Practices*

**Guy B. Sereff**

19 March 2014

# About The Presenter

**Guy B. Sereff**

- Author, Speaker and Technology Practitioner
- Vice President / Enterprise Architecture
- Technology Industry Experience
  - *Application Research & Development (12 years)*
  - *Large-Scale Technology Management (8 years)*
  - *Global Enterprise Architecture (8 years)*
- Enterprise Architecture Domain Experience
  - *Business Architecture*
  - *Information Architecture*
  - *Application Architecture*
  - *Solution Architecture*
  - *Architecture Governance*
- Pragmatic Blend of Strategy and Tactical Execution

http://www.linkedin.com/in/guysereff

# Agenda

**Basic Big Data Concepts and Characteristics**

- Early Origins
- Working Definition(s)
- Characteristics
- Enabling Technologies

**The Changing Landscape of Analytics**

**Best EA Practices for Big Data**

- Establish Adaptive Enterprise Data Principles
- Don’t Abandon the Enterprise Data Model or Data Governance
- Establish a Big Data Reference Architecture
- Clarify Big Data Roles, Accountabilities and Decision Rights

**Recommended Next Steps**

**Questions and Closing Comments**

# Early Conceptual Origins of "Big Data"

**1997**: NASA identified 'big data' as an emerging problem when trying to visualize computational fluid dynamics

**2000**: Commercial considerations were noted at the Eighth World Congress of the Econometric Society

**2001**: The META Group documented what has become the widely accepted *'3V's of Big Data'*

"Visualization provides an interesting challenge for computer systems: data sets are generally quite large, taxing the capacities of main memory, local disk, and even remote disk. We call this the problem of **big data**."

Cox and Ellsworth (1997). IEEE Visualization Conference

"**Big Data** refers to the explosion in the quantity (and sometimes quality) of available and potentially relevant data, largely the result of recent and unprecedented advancements in data recording and storage technology. In this new and exciting world, sample sizes are no longer fruitfully measured in "number of observations," but rather in, say megabytes"

Diebold (2000). World Congress of the Economics Society

"While enterprises struggle to consolidate systems and collapse redundant databases to enable greater operational, analytical, and collaborative consistencies, changing economic conditions have made this job more difficult. E-commerce, in particular, has exploded data management challenges along three dimensions: **volumes, velocity, and variety**. "

Laney (2001). The META Group

# Big Data – Working Definitions

## Gartner

Big Data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision-making.

## InfoWorld

In the never-ending quest for a competitive advantage, organizations are turning to large repositories of corporate and external data to uncover trends, statistics, and other actionable information to help decide on their next move. Those data sets, along with their associated tools, platforms, and analytics, are often referred to as *big data*.

## Forrester

[Pragmatic] Big Data is the frontier of a firm's ability to store, process, and access (SPA) all the data it needs to operate effectively, make decisions, reduce risks, and serve customers.

## Mayer-Schönberger & Cukier

"There is no rigorous definition of Big Data.

...Big Data refers to things one can do at a large scale that cannot be done at a smaller one, to extract new insights or create new forms of value in ways that change markets, organizations, the relationship between citizens and government and more."

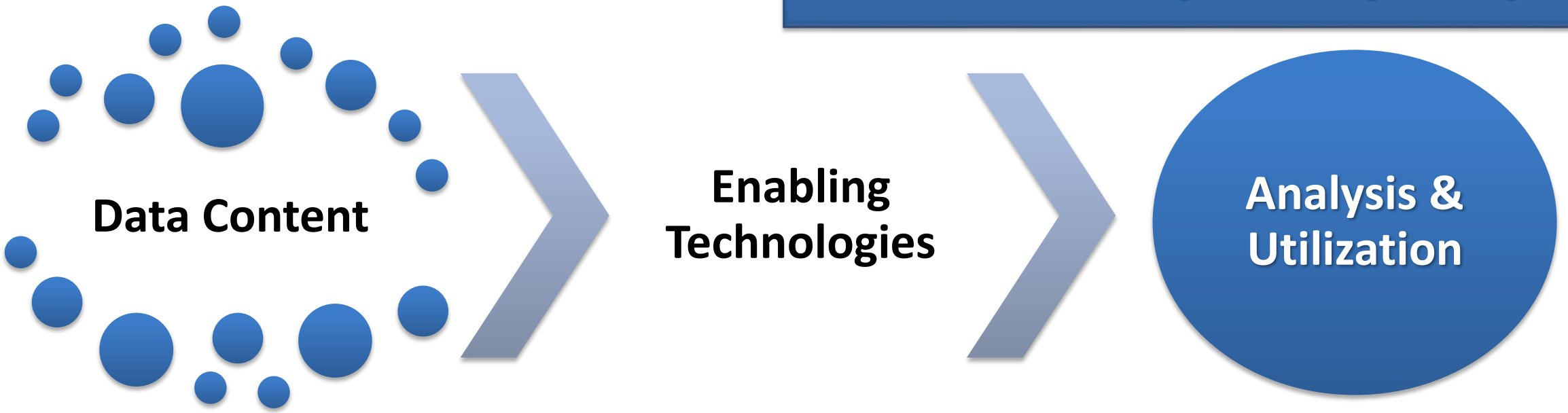
# Common Aspects of Big Data

Indiscriminant use of the term *Big Data* is common

"Big Data" generally covers three aspects

- Data Content
- Enabling Technologies
- Analysis and Utilization

- **Big Data Content**: Massive amounts of data needing to be processed with the common characteristics of *volume, velocity and variety*
- **Big Data Enabling Technologies:** Tools and techniques capable of processing big data content
- **Big Data Analytics**: Analysis, interpretation and actions based on results from big data enabling technologies



# Data Characteristics – *The 3V's of Big Data*

**Volume**

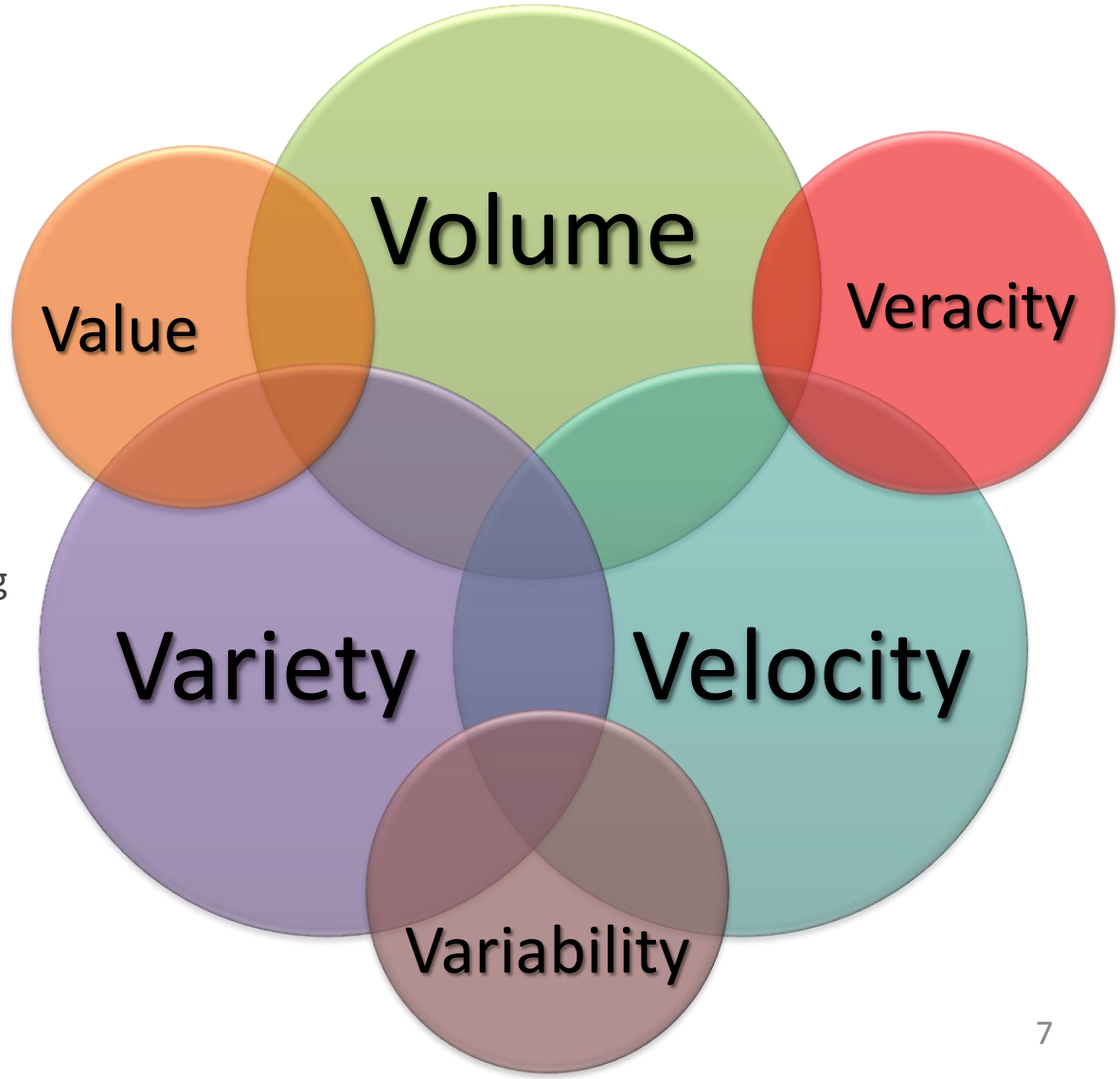
- Sheer amount of data organizations are amassing in their systems of record, data stores and data warehouses
- World's information doubles every two years; IDC anticipates 40 *Zettabytes* of digital data by 2020 *(≅ 43 Billion Terabytes)*

**Velocity**

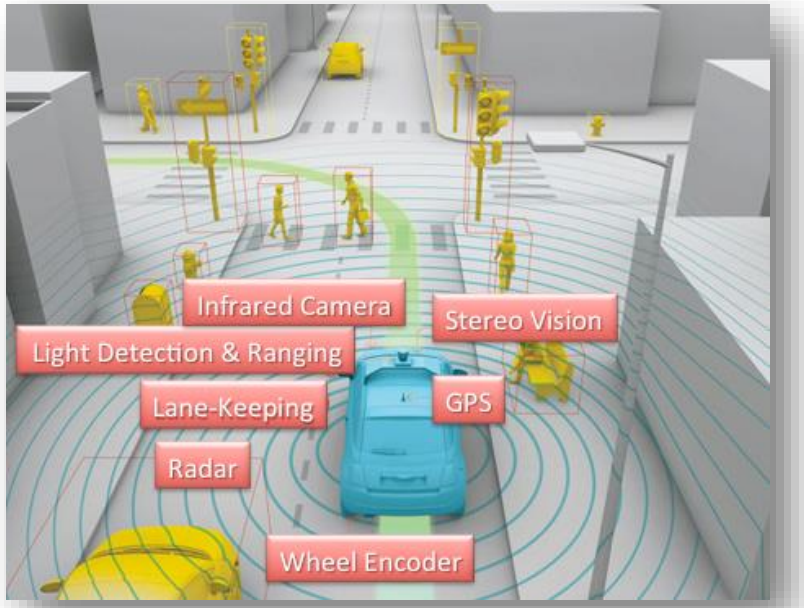
- The pace at which data is generated or received, as well as how timely the assessment of the data needs to be
- Incoming data often must be analyzed faster than it can be stored and subsequently retrieved for critical decision-making

**Variety**

- Growing use of unstructured data, such as Email, customer service transcripts, blogs, forums and video
- Increase use of external data sources, such as social media outlets, industry research data, and market research data



# The 3Vs in Action - Autonomous Car Technology




**Volume**: 100's to 1,000's of Sensors/Data Capture Points

**Velocity**: Mission Critical Real Time Decisioning

**Variety**: Wide Assortment of Data Types

Now consider the complexity of managing a metropolitan area freeway system filled with hundreds of thousands of *human operated* and *driverless* cars interacting on the same roadways!!

...then add random events outside the control of the 'system'...

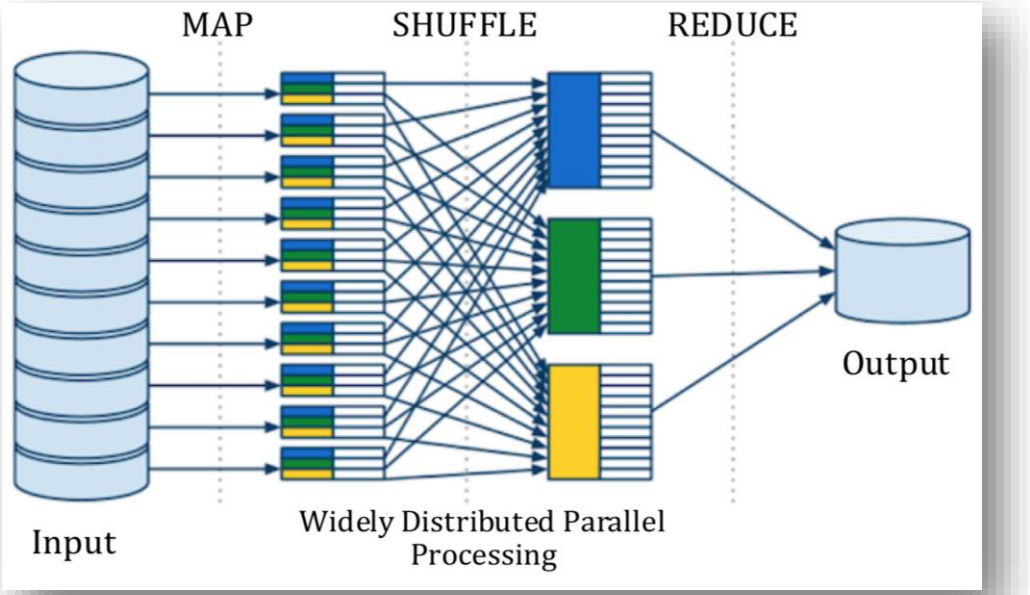
# Enabling Technologies – Map Reduction (MapReduce)

Developed at Google for indexing web pages, replacing their original indexing algorithms and heuristics in 2004

Software framework to process massive amounts of unstructured data in parallel across a distributed cluster of processors or stand-alone computers

Uses nominal commodity processing nodes versus traditional EDW dedicated data appliances, some portion of which are anticipated to fail during operation

Successfully being used against petabytes of input data being processed across thousands of independent nodes




**Map Reduction Process**

- **Map** very large data sets into meaningful key/value pairs across parallel processing nodes
- **Shuffle** data into common key groupings
- **Reduce** the shuffled data into a manageable aggregation result set

# Enabling Technologies - NoSQL

"NoSQL is a non-relational database that stores and accesses data using key-values. Instead of storing data in rows and columns like a traditional database, a NoSQL DBMS stores each item individually with a unique key. Additionally, a NoSQL database does not require a structured schema that defines each table and the related columns. This provides a much more flexible approach to storing data than a relational database."

Techterms.com (2014)

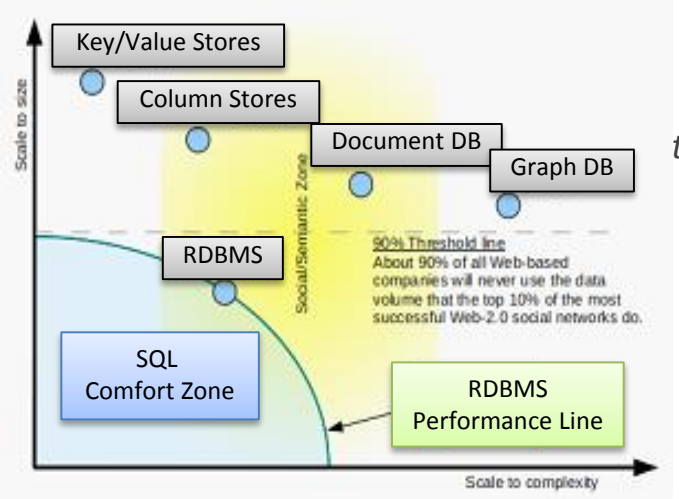
## Common NoSQL Database Types

**Key-Value Stores**
*Use a hash table of key/pc*

**Column Stores**
*Use key/column pairs, whe* *ted across multiple servers*

**Document Databases**
*Information stored as vers collections of key/value co*

**Graph Databases**
*A data model based on gr rather than row, columns*




| RDBMS/SQL | NoSQL |
|---|---|
| • Data Stored in Columns and Tables<br>• Relationships Represented by Data<br>• Data Manipulation Language<br>• Data Definition Language<br>• ACID Transactions *(Atomic, Consistent, Isolated, Durable)*<br>• Abstraction from Physical Layer<br>• Pre-Defined Schema<br>• Well-Defined Semantics<br>• Standard Definitions | • Large Data Volumes<br>• Scalable Replication and Distribution<br>• Rapid Query Response Needed<br>• Mostly Queries, Few Updates<br>• Asynchronous Inserts/Updates<br>• Schema-less<br>• BASE Transaction *(Basically Available, Soft state, Eventually consistent)*<br>• CAP Theorem<br>• Open Source Development |

# Enabling Technologies - Hadoop

Apache Software Foundation open source framework for reliable, scalable distributed computing

- Started in 2005; Version 2.3.0 released Feb 2014
- Four Core Modules
- Ten additional Hadoop-Related projects underway

Several leading technology vendors offer Apache Hadoop as part of their product lines, including IBM, Intel, Amazon, VMware and Hortonworks

Leading organizations are actively using Hadoop including Adobe, Alibaba, eBay, Facebook, Google, Hulu, IBM, LinkedIn, Twitter, Yahoo! and many more

| Core Hadoop Modules | |
|---|---|
| **Hadoop Common**<br>*The common utilities that support the other Hadoop modules* | **HDFS™**<br>*A distributed file system that provides high-throughput access to application data* |
| **Hadoop Yarn**<br>*A framework for job scheduling and cluster resource management* | **Hadoop MapReduce**<br>*A YARN-based system for parallel processing of large data sets* |

**Hadoop-Related Projects**

**Ambari™** - A web-based tool for provisioning, managing, and monitoring Apache Hadoop clusters

**Hive™** - A data warehouse infrastructure that provides data summarization and ad hoc querying

**Avro™** - A data serialization system

**Mahout™** - A scalable machine learning and data-mining library

**Cassandra™** - A scalable multi-master database with no single point of failure

**Pig™** - A high-level data-flow language and execution framework for parallel computation

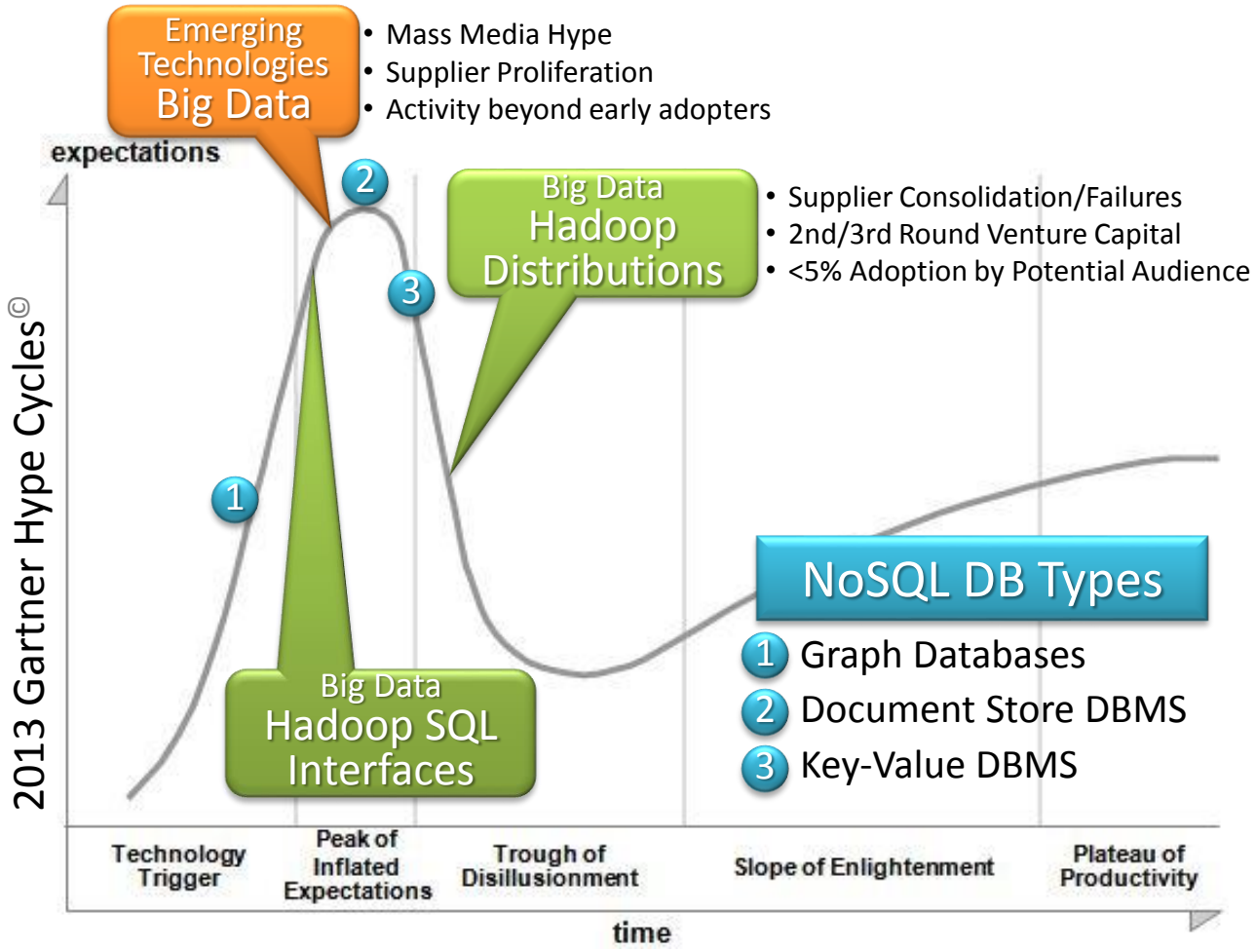
**Chukwa™** - A data collection system for managing large distributed systems

**Spark™** - A fast and general compute engine for Hadoop data

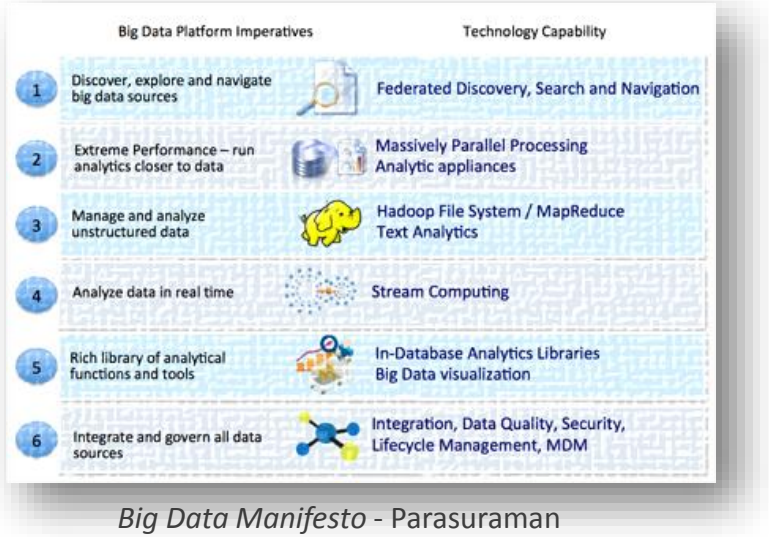
**HBase™** - A scalable, distributed database that supports structured data storage for large tables

**ZooKeeper™** - A high-performance coordination service for distributed applications

# Enabling Technologies – Hadoop Trends




In 2012 Forrester Research identified Hadoop as "the nucleus of the next-generation EDW [Enterprise Data Warehouse] in the cloud", but cautioned Big Data practitioners about the risks stemming from the relatively low level of maturity of enterprise-grade Hadoop offerings at the time.




*Big Data Manifesto* - Parasuraman

# The Changing Landscape of Analytics

***So what benefit do we get from all of this?***
The real potential value found within the massive amounts of data is the insight we are able to gain from the use of Big Data, analyzing and deriving conclusions that were not otherwise discernable from traditional data viewpoints.

***Business Intelligence*** provides transactional trend analysis to hypothesize about future activity, given a list of variables and correlational data points

Challenge using platforms based on structured and semi-structured data to process unstructured data

Volume and Velocity aspects of Big Data typically out-strip traditional processing environments

Need the ability to support rapid 'what-if' scenario iterations

| Traditional Data Warehouse |
| --- |
| Complete Record from Transaction System |
| All Data Centralized |
| Addition Every Month/Day of New Data |
| Analytics Designed Against Stable Environment |
| Many Reports Run on a Production Basis |

# The Changing Landscape of Analytics

| | Structured | Hybrid | Content |
|---|---|---|---|
| **Descriptive** *What is happening?* | *Who are my top 10 customers and products?* | *Who are my top 10 influencers?* | *What are my top 10 product defects?* |
| **Diagnostic** *Why did it happen?* | *What is the source of declining sales?* | *What is the root cause of poor sentiments?* | *Why is my sentiment worse than my competitors?* |
| **Predictive** *What will happen?* | *What is likely to be a fraudulent transaction?* | *Which patient is likely to enter a hospital in a year?* | *Is this package likely to be suspicious?* |
| **Prescriptive** *What should I do?* | *Which as should I present based on profile and transactions?* | *Which ad should I present based on profile influence sentiments, video and transactions?* | *Which ad should I present based on video?* |

*Comparison of Data Structure Use Cases* - Elliot

Traditional environments and data management tools are capable of handling the majority of an institution's current data needs

Dedicated data warehouse hardware appliances, Massively Parallel Processing (MPP) techniques and common SQL instructions are successfully processing massive amounts of structured data

Think *Both/And* versus *Either/Or* - traditional data warehouse *and* Big Data analytic environments each have appropriate use cases within the enterprise and are complimentary platforms

Acceptance of imprecision inherent in loosely structured data requires new levels of data accuracy tolerance thresholds

*Fun Fact*: E. F. Codd faced a lot of opposition to his 'radical' relational database model 40+ years ago

# The Changing Landscape of Analytics

| | Structured | Hybrid | Content |
|---|---|---|---|
| **Descriptive** *What happened?* | *Who are my top* | *Who are my top* | *What are my* |
| **Diag** *Why happ* | | | |
| **Pred** *Wha happ* | | | |
| **Pres** *Wha do?* | *and transactions?* | *influence sentiments, video and transactions?* | |

**Top 8 Emerging Big Data Career Opportunities**

- ETL Developers
- Hadoop Developers
- Visualization Tool Developers
- Data Scientists
- OLAP Developers
- Data Warehouse Appliance Specialist
- Predictive Analytics Developers
- Information Architects

*CIO Magazine – Jan 2014*

*Comparison of Data Structure Use Cases* - Elliot

Traditional environments and data management tools are capable of handling the majority of an institution's current data needs

Dedicated data warehouse hardware appliances, Massively Parallel Processing (MPP) techniques and common SQL instructions are successfully processing massive amounts of structured data

Think *Both/And* versus *Either/Or* - traditional data warehouse *and* Big Data analytic environments each have appropriate use cases within the enterprise and are complimentary platforms

Acceptance of imprecision inherent in loosely structured data requires new levels of data accuracy tolerance thresholds

*Fun Fact*: E. F. Codd faced a lot of opposition to his 'radical' relational database model 40+ years ago

# The Changing Landscape of Analytics

***4 Ways to Actually Use Big Data* – 2014 Inc. Magazine**

| | |
|---|---|
| **Use filters to separate important messages from background noise** | • Dell Computers receives more than 25,000 social media mentions daily in 11 different languages<br>• Dell automatically filters for messages that could go viral if unaddressed<br>• *Filters reduce the flood of social media data to a manageable stream* |
| **Track overall changes in message volume** | • Tweets about British Airways spiked abnormally due to a disgruntled passenger who tweeted to 50,000 users in the US and UK of his lost luggage<br>• BA took 10 hours to respond – by then the story spread to news outlets on both continents<br>• *Track changes in message volume on social media to head off PR nightmares and capitalize on positive trends* |
| **Incorporate tools that automatically track sentiment** | • Domino's Pizza employee posted a viral video of himself, excavating his nasal cavity at work<br>• One million views later, news outlets were carrying the story and customer ratings plummeted<br>• *Instantly track changes in sentiment through Social Analytics Software* |
| **Choose software that spits out snazzy reports** | • Southwest Airlines has a team focused on their social media community of 6M followers<br>• Executive social media skepticism persists, seeing Twitter and Facebook as "soft" network tools<br>• *Analytic tools offer sophisticated business reporting functions to track changes in visibility and sentiment over time and compared to competitors* |

# Best Enterprise Architecture Practices for Big Data

Big Data represents a large investment and offers new insights to fuel opportunities across the enterprise

Architects should consider the following best practices for Big Data

- Establish Adaptive Enterprise Data Principles
- Don't Abandon the Enterprise Data Model or Data Governance
- Establish a Big Data Reference Architecture
- Clarify Big Data Roles, Accountabilities and Decision Rights

# Best Enterprise Architecture Practices for Big Data

Big Data represents a large investment and offers new insights to fuel opportunities across the enterprise

Architects should consider the following best practices for Big Data

- Establish Adaptive Enterprise Data Principles
- Don't Abandon the Enterprise Data Model or Data Governance
- Establish a Big Data Reference Architecture
- Clarify Big Data Roles, Accountabilities and Decision Rights

# Establish Adaptive Enterprise Data Principles

Big Data challenges traditional Enterprise Data Principles

- Lack of traditional schema definitions
- Unstructured nature
- Use of potentially unsecure external sources

Enterprise Architects should work closely with Information and Data Architects to revisit and refactor the existing principles

- At some point Big Data and 'regular' data will intersect
- Need to ensure loosely structured data does not taint or invalidate critical transactional data
- Ensure regulatory constraints are met

Each organization will have to determine the appropriate level of adaptation

- Ensure principles reflect the needs of both traditional data and Big Data
- Don't assume this will be the last paradigm shift – establish an adaptive, evergreen approach

**TOGAF 9 Data Principles**

- Data is an asset
- Data is shared
- Data is accessible
- Data has a trustee
- Data has common vocabulary/data definitions
- Data should be secure

# Don’t Abandon the Enterprise Data Model or Data Governance

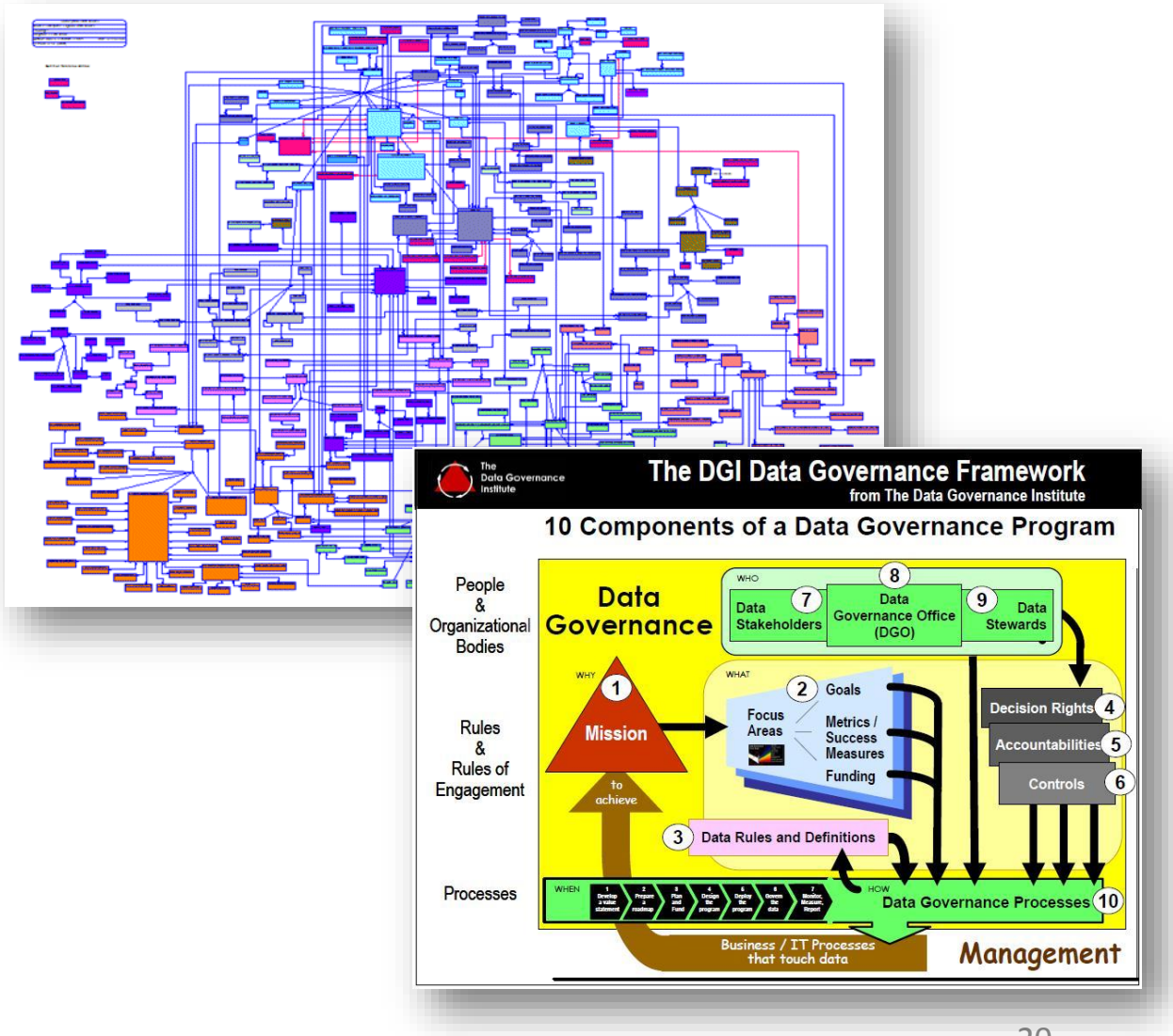
Big Data seems diametrically opposed to traditional controls

- Unstructured data hard to map to the Enterprise Conceptual Data Model (ECDM)
- External sources of data lack similar controls
- Off-Premise Cloud Nodes provide added security risk
- Often introduced outside the purview of EA

Collision of Big Data with the rest of the enterprise data landscape is inevitable

Enterprise Architects, Information Architects, Data Architects and Data Engineers should work together

- Establish an appropriate level of governance that is congruent with the broader Enterprise Data Governance policy, while accommodating the unique aspects of Big Data to help its rapid progression across the organization
- Map and gap Big Data content to the ECDM, refactoring the model as needed



# Don’t Abandon the Enterprise Data Model or Data Governance

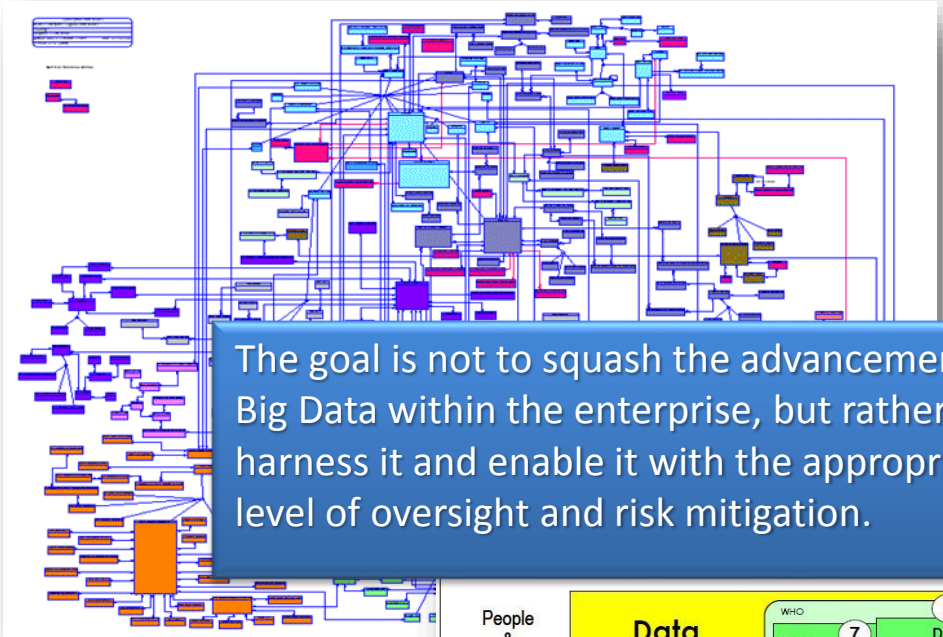
Big Data seems diametrically opposed to traditional controls

- Unstructured data hard to map to the Enterprise Conceptual Data Model (ECDM)
- External sources of data lack similar controls
- Off-Premise Cloud Nodes provide added security risk
- Often introduced outside the purview of EA

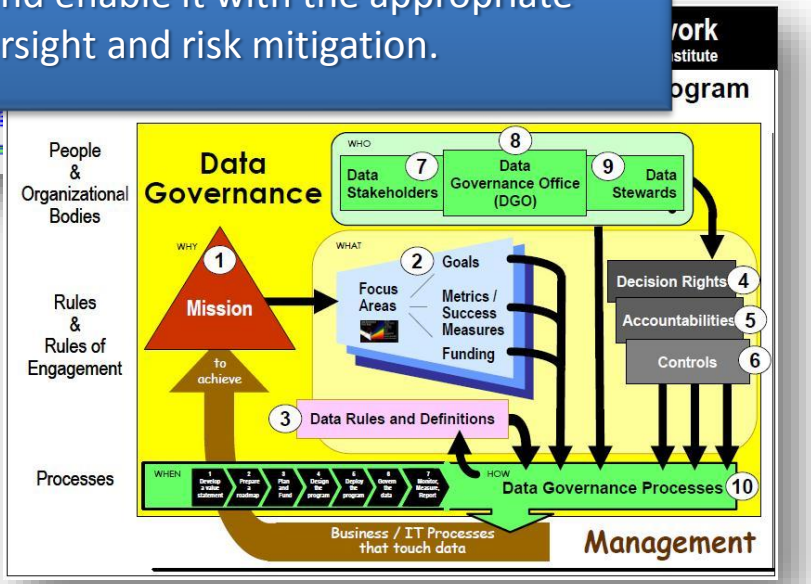
Collision of Big Data with the rest of the enterprise data landscape is inevitable

Enterprise Architects, Information Architects, Data Architects and Data Engineers should work together

- Establish an appropriate level of governance that is congruent with the broader Enterprise Data Governance policy, while accommodating the unique aspects of Big Data to help its rapid progression across the organization
- Map and gap Big Data content to the ECDM, refactoring the model as needed



The goal is not to squash the advancements of Big Data within the enterprise, but rather harness it and enable it with the appropriate level of oversight and risk mitigation.



# Establish a Big Data Reference Architecture

Most organizations have found ways to safely introduce new technology into the corporate ecosystem with appropriate controls and agility

Big Data seems to find its way into organizations through both formal and informal paths, creating multiple *de facto* 'standard' Big Data solutions

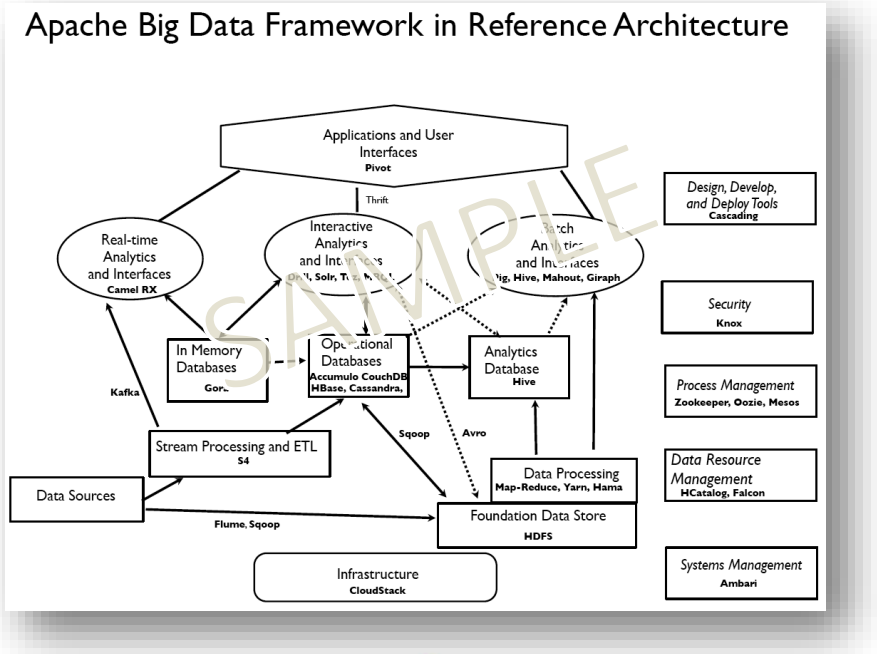
It is imperative to establish an Enterprise Big Data Reference Architecture

- Align with current reference architecture practice
- Enable tool selection and rapid environment provisioning
- Establish best practices and implementation patterns
- Identify competing solutions and create an aggressive roadmap for alignment

**Reference Architecture**

A predefined architectural pattern, or set of patterns, possibly partially or completely instantiated, designed, and proven for use in particular business and technical contexts, together with supporting artifacts to enable their use. Often, these artifacts are harvested from previous projects.

*- Reference Architecture: The Best of Best Practices - IBM*



# Clarify Big Data Roles, Accountabilities and Decision Rights

Assemble a comprehensive roster of current and prospective Big Data participants, along with a complete set of user stories and use cases

Align roles/tasks to other data-centric disciplines such as EDW, BI and Data Analytics

Gain consensus among Big Data ecosystem constituents, including suppliers and consumers

Once ratified, ensure that Big Data roles, accountabilities and decision rights are widely communicated and adhered to

Be willing to slow the Big Data team down long enough to help them get organized for longer term success

Each firm must determine for themselves the correct level of decision federation consistent with their prevailing risk management practices

**Decision Rights – An Example**

- Who gets to decide on tools?
- Who selects the commodity servers used for MapReduce operations?
- Who determines which external sources of unstructured data are acceptable?
- Who gets to map Big Data information to the rest of the data environment?
- Who approves deviation requests?
- Who determines the appropriate level of confidence required for various investment levels?

# Recommended Next Steps

Recognize that Big Data is more than a passing techno-trend and will need to co-exist with traditional data

Monitor domain maturity and align roadmap accordingly

Ensure Enterprise Architecture function is fully engaged in Big Data practices across the organization

Provide flexibility to accommodate the unique needs of Big Data without compromising the broader needs of the organization

Monitor for potential benefits:

- Insightful and actionable analytics based on new and unique data sources
- Alignment to and consistency with the organization's data practices
- Repeatable investment criteria and deployment patterns across the Big Data spectrum
- Better internal collaboration across the Big Data communities
- Successful cohabitation of Big Data with existing data resources

**Superior Commanders succeed in situations where ordinary people fail because they obtain more timely information and use it more quickly.**

*-Sun Tsu, The Art of War, 6th Century BC*

# Any Questions?

Orbus Software Group

@OrbusSoftware

Download this presentation and accompanying white paper from:
**www.orbussoftware.com/downloads**