Big Data Systems and Interoperability

Emerging Standards for Systems Engineering

David Boyd
VP, Data Solutions
Email: dboyd@incadencecorp.com

InCadence
STRATEGIC SOLUTIONS

Topics

- Shameless plugs and denials
- What is Big Data and Why do we do it?
- Standards Evolution Timeline
- NIST Big Data Interoperability Framework
- ISO/IEC JTC1 WG9 Big Data
- Future Standard Evolution
- Conclusion

InCadence STRATEGIC SOLUTIONS

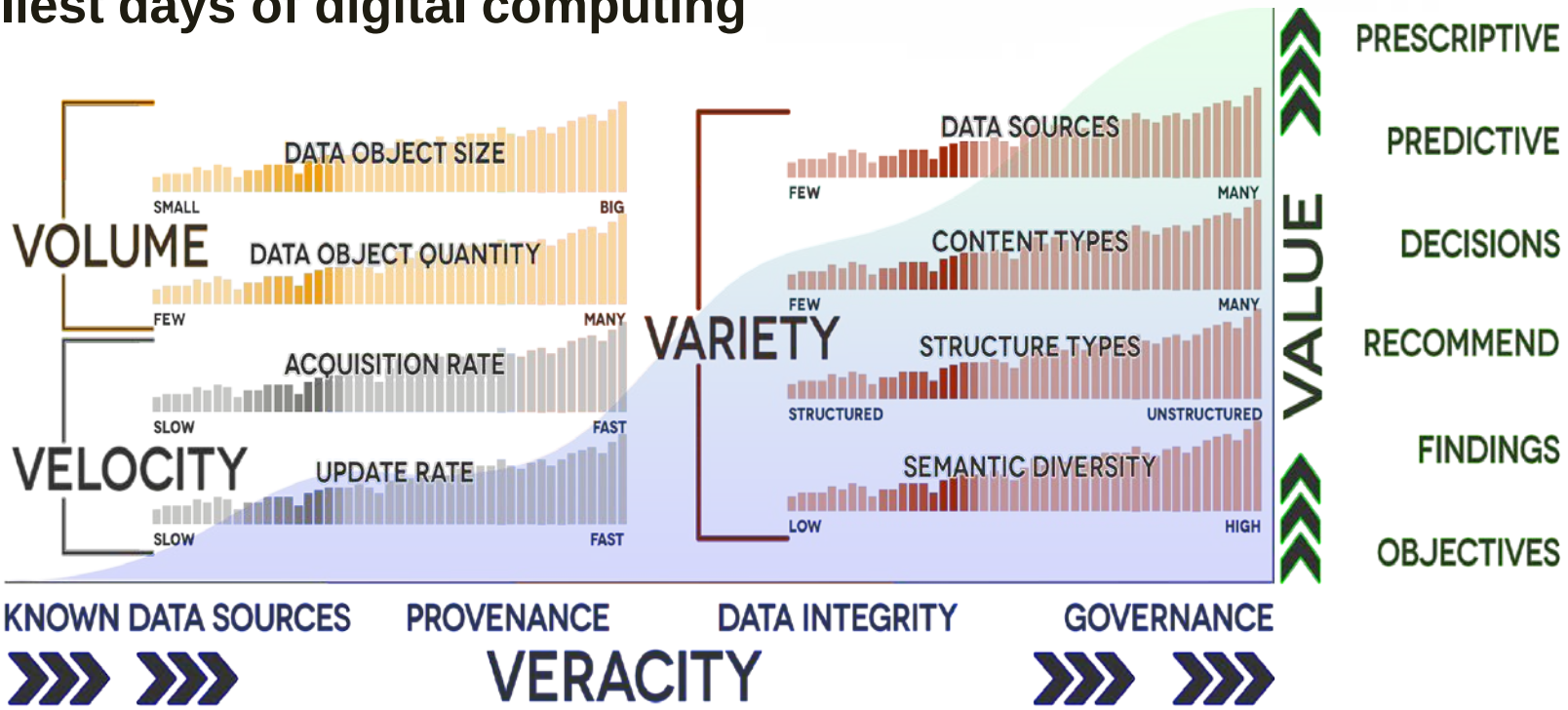
Shameless Plugs & Denials

- InCadence Strategic Solutions
  - Woman Owned Small Business founded in 2009
  - 90+ people, >80% with DoD clearances
  - Customers include Army, Department of State, FBI, and Navy
  - Deep Institutional Expertise in:
    - Biometrics
    - Data Architectures (e.g. FEA DRM)
    - Data Interchange (e.g. NIEM)
    - Big Data Solutions (e.g. Haddoop, NoSQL, SPARK, etc.)
- My self
  - 35 years of software development, systems integration and system architectures for data intensive system
  - Supported Big Data standards since 2013
    - Editor on NIST and ISO/IEC documents
    - Chair INCITS Technical Committee on Big Data
  - Robotics Mentor and 3D Printing addict
- The views expressed in this presentation are mine and mine alone. Any semblance to the views of NIST, INCITS, or ISO is purely coincidental

InCadence STRATEGIC SOLUTIONS

What is Big Data

- **Gartner – 3Vs (Volume, Velocity, Variety) & additional Vs cropping up (Veracity, Value, Variability)**
- **Big is a relative term – we have been dealing with the Vs since the earliest days of digital computing**

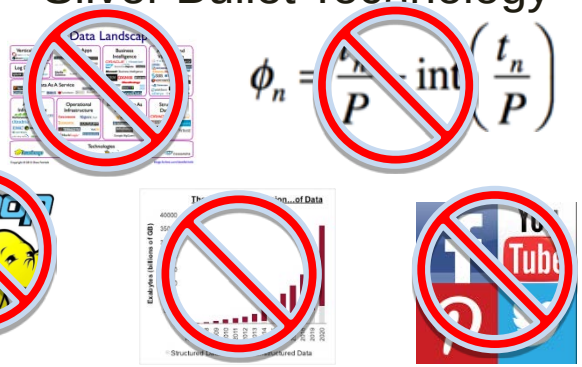


The real issue is not the data - it is the problems we are seeking to answer with the data - Big "Data Problems"

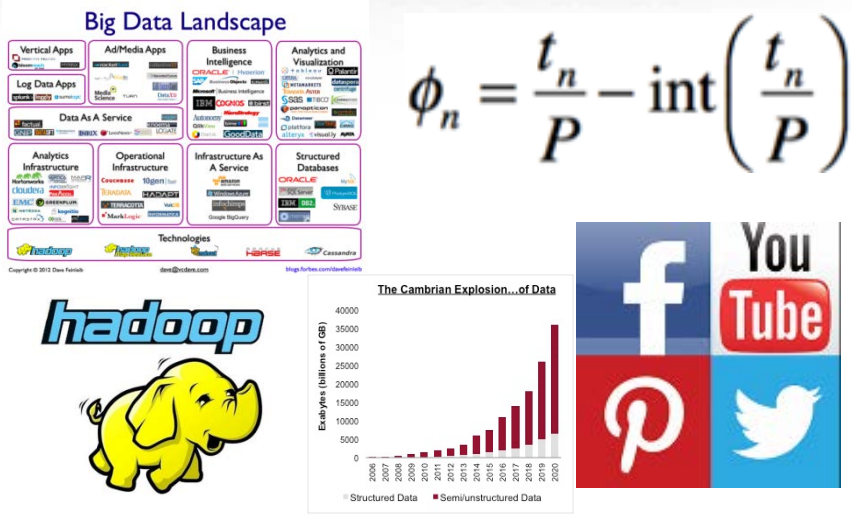
InCadence STRATEGIC SOLUTIONS

Why do Big Data?

- NOT SINGULARLY about
  - Large volumes of data
  - Social Media Data
  - Data Science/Analytics
  - HADOOP
  - Graphics and Charts
  - Silver Bullet Technology



Is COLLECTIVELY about leveraging ALL of them to produce VALUE from DATA



"The goal is to turn data into information, and information into insight."
Carly Fiorina, former CEO of Hewlett-Packard

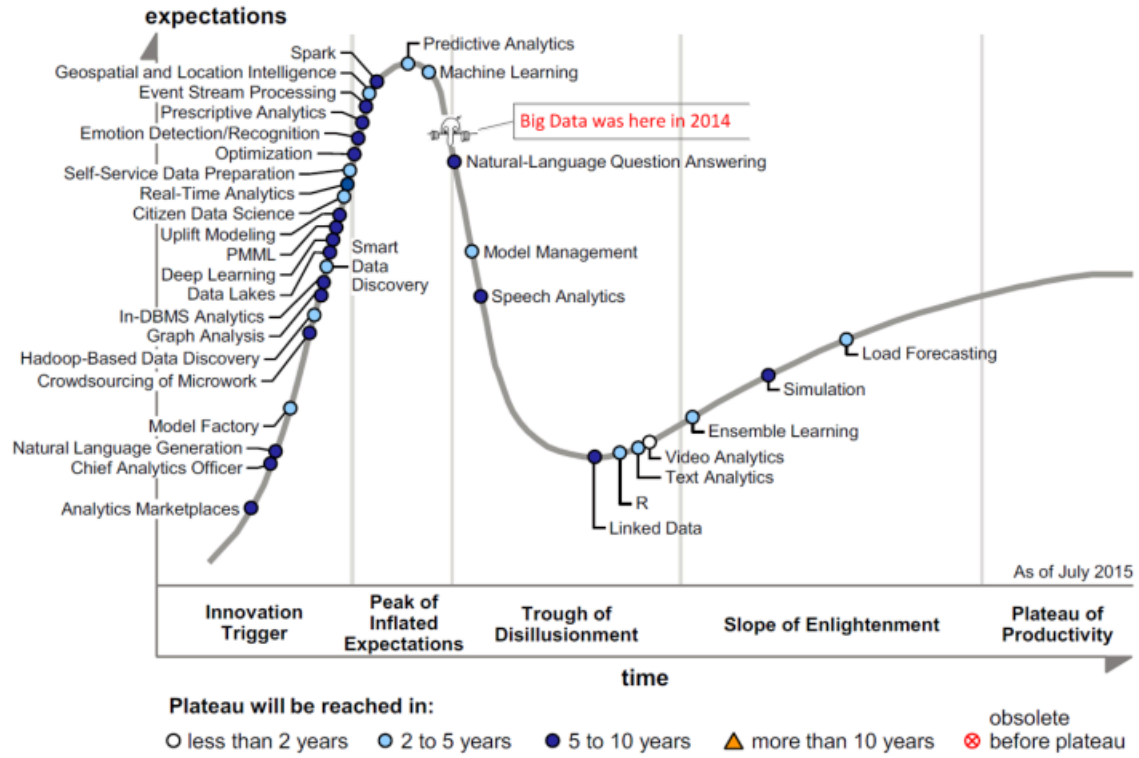
65% of organizations felt they were effective at capturing data, but just 46% were effective at disseminating information and insights.
MIT Sloan Mgmt Review

The goal of Big Data is to derive value that leads to intelligent decision making

InCadence STRATEGIC SOLUTIONS

Why Big Data Standards

Figure 1. Hype Cycle for Advanced Analytics and Data Science, 2015



Source: Gartner (July 2015)

Because analytics using multiple data sources and structures has become the norm, information architects must focus on adapting to new data quickly, and on coherently managing diverse information and analytics - Gartner

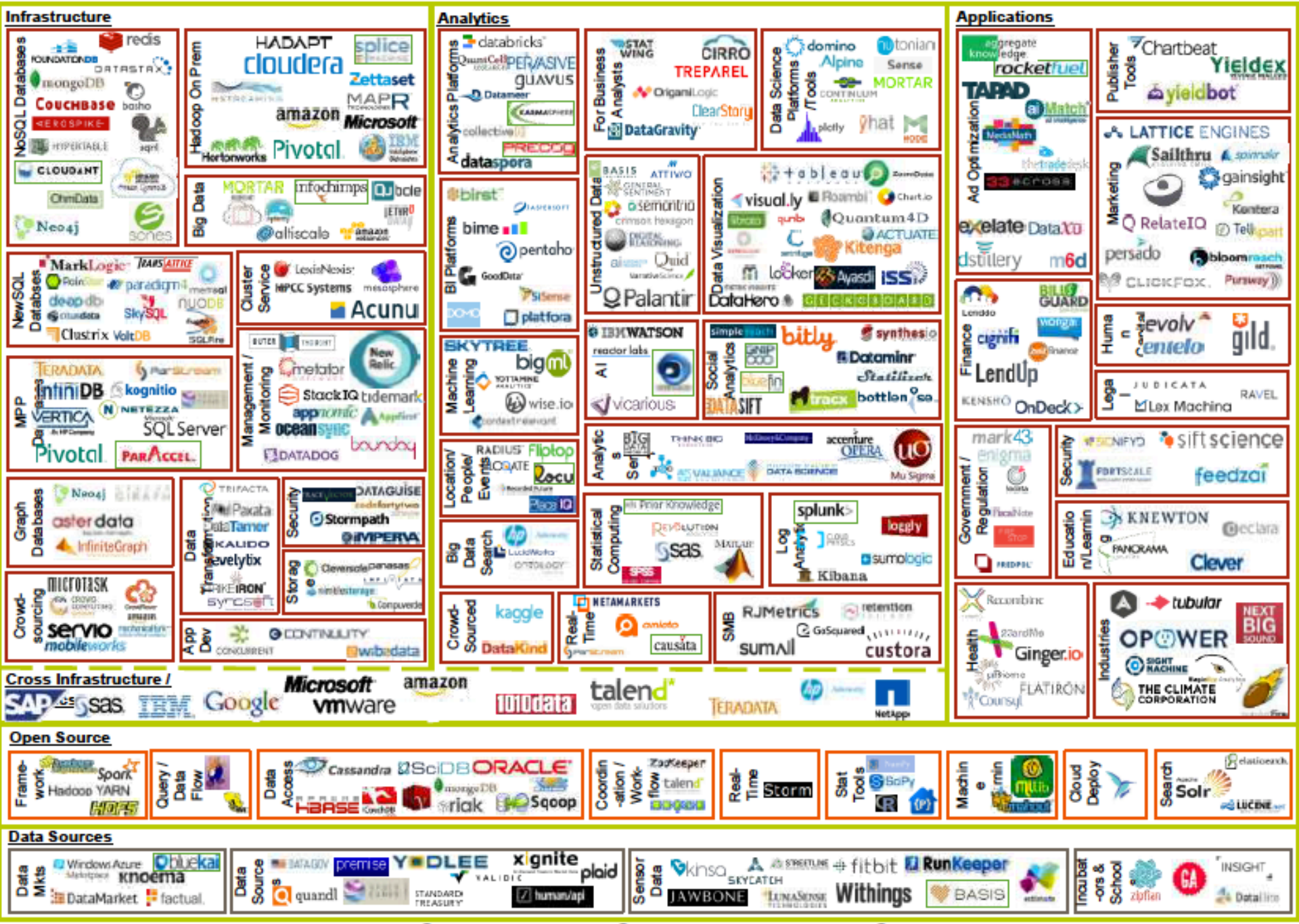
"... is now being replaced by practicality, because the technology and information asset types offer new alternatives that are most often additive or complementary to long-standing, traditional practices." - Gartner

- ✧ The Hype is over – we need to move to productivity
- ✧ The Ecosystem of tools and technologies continues to expand
- ✧ Some how all of this needs to work together

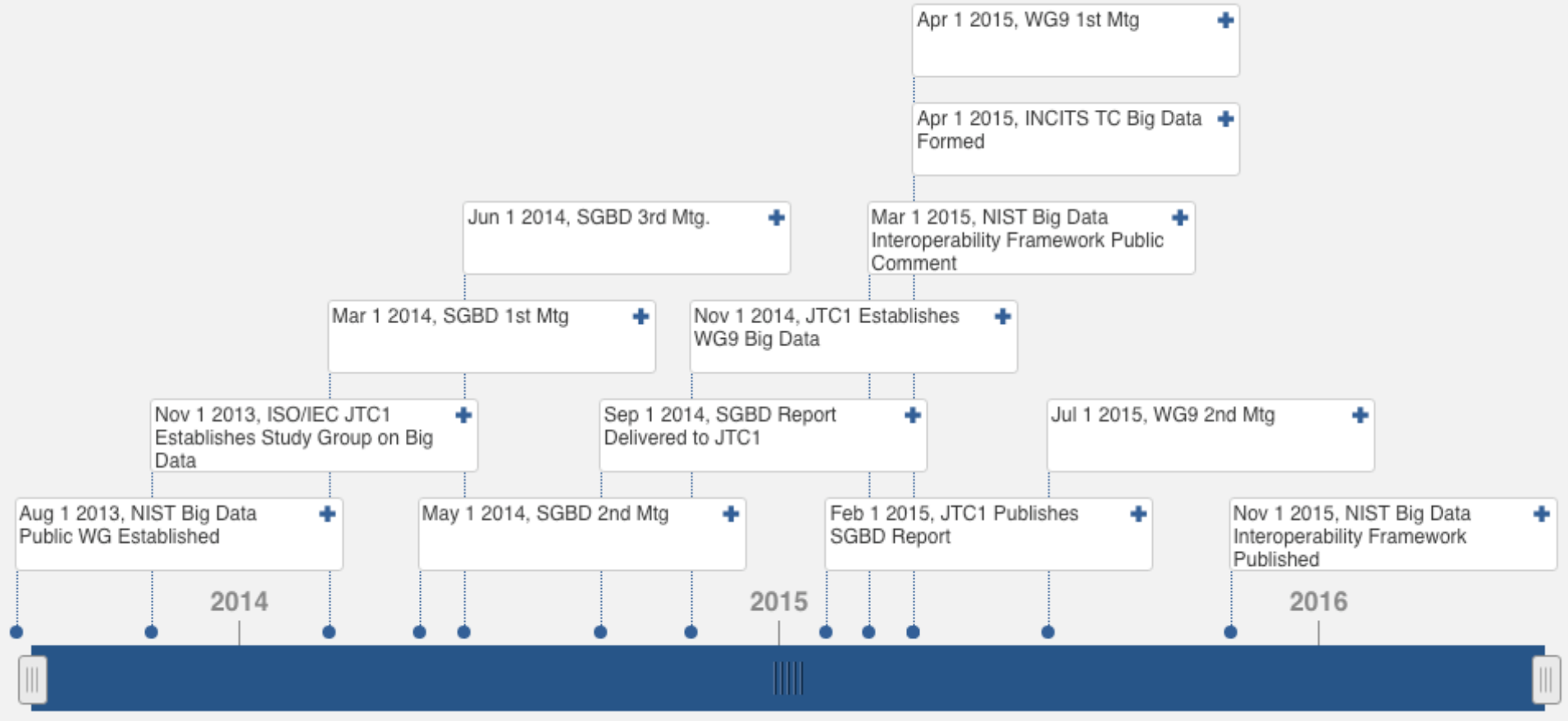
InCadence STRATEGIC SOLUTIONS

BIG DATA LANDSCAPE, VERSION 3.0

Exited: Acquisition or IPO

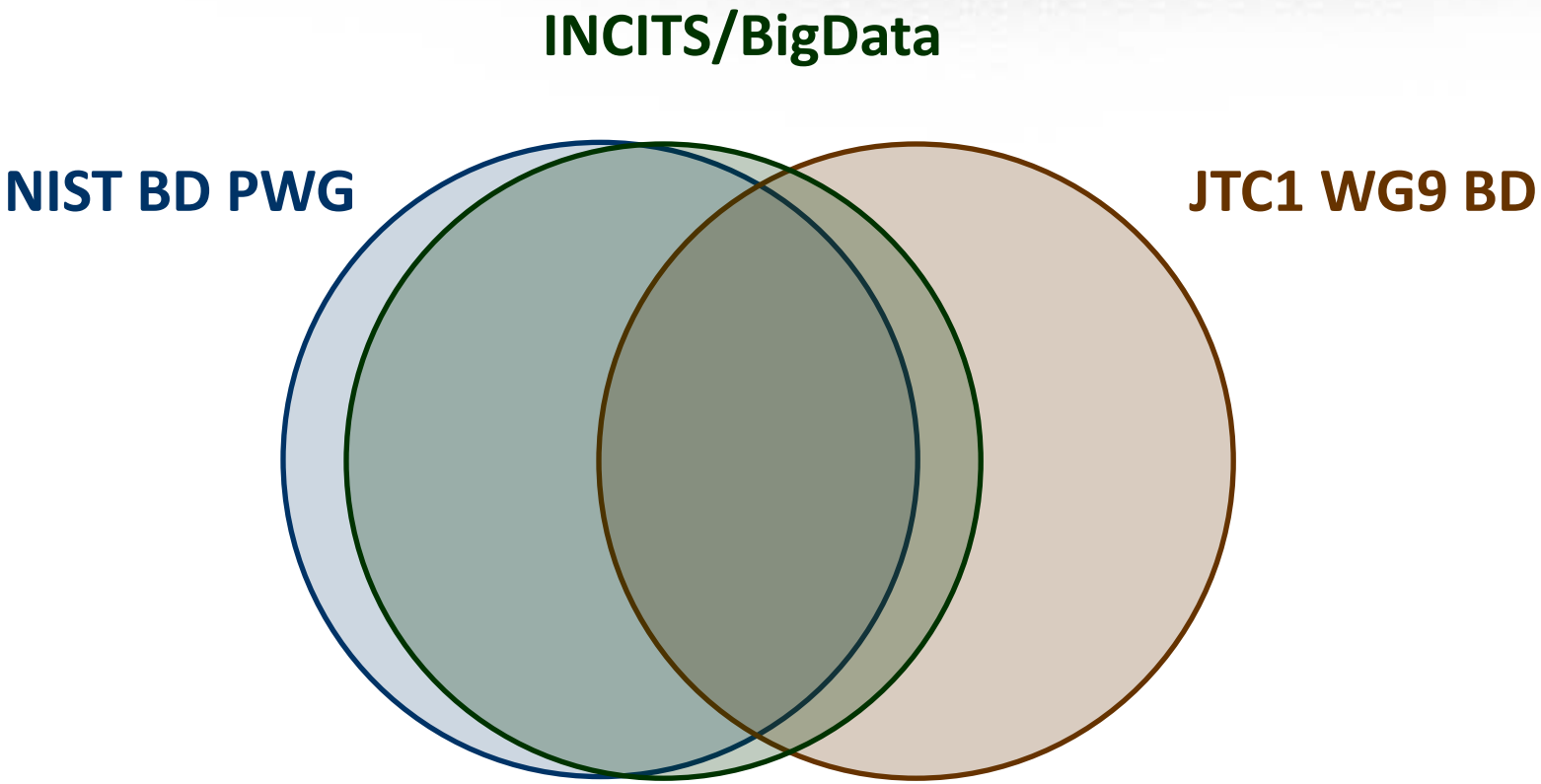


Big Data Standards Timeline (Historic)



InCadence STRATEGIC SOLUTIONS

Overlapping Membership on Efforts



Overlaps are indicative, but not to scale.

InCadence STRATEGIC SOLUTIONS

NIST Big Data Interoperability Framework

- Seven Volumes
  - Volume 1, Definitions
  - Volume 2, Taxonomies
  - Volume 3, Use Cases and General Requirements
  - Volume 4, Security and Privacy
  - Volume 5, Architectures White Paper Survey
  - Volume 6, Reference Architecture
  - Volume 7, Standards Roadmap
- Latest versions available: http://bigdatawg.nist.gov/V1_output_docs.php
- Published October 2015 as NIST SP 1500-n

InCadence STRATEGIC SOLUTIONS

Volumes 1: Definitions

- Define a common vocabulary for multiple audiences
- Set the landscape and issues around big data
  - Defined a number of related terms
- Two key aspects
  - Focused on Characteristics (the Vs)
  - Focused on need for scalable architectures
- Issues
  - Definitions need to be more normative

Big Data *consists of extensive datasets—primarily in the characteristics of volume, variety, velocity, and/or variability—that require a scalable architecture for efficient storage, manipulation, and analysis.*

InCadence STRATEGIC SOLUTIONS

Volumes 2: Taxonomies

- Define actors and roles as used within the reference architecture
- Start to define data characteristics
- Issues
  - Our eyes were bigger than our stomach – How do you not do a taxonomy of all computing?

InCadence STRATEGIC SOLUTIONS

Volume 3: Big Data Use Cases and Requirements

- Built from responses to Use Case survey (26 fields)
- 51 Responses
  - Government Operations (4)
  - Commercial (8)
  - Defense (3)
  - Healthcare and Life Sciences (10)
  - Deep Learning and Social Media (6)
  - The Ecosystem for Research (4)
  - Astronomy and Physics (5)
  - Earth, Environmental and Polar Science (10)
  - Energy (1)
- Decomposed then aggregated into 34 general requirements across 6 categories
  - Data Source Requirements (3)
  - Transformation Provider Requirements (3)
  - Data Consumer Requirements(6)
  - Security and Privacy Requirements (2)
  - Lifecycle Management Requirements (9)
  - Other Requirements (5)
- Detailed requirements all traceable to general requirements
- Issues
  - We didn't know what we didn't know – template was overly simplistic
  - Additional use cases needed

InCadence STRATEGIC SOLUTIONS

Volume 4: Security and Privacy

- Recognized early on as a key concern requiring a more complete treatement
- Describes:
  - S&P issues particular to Big Data
  - Some S&P specific use cases
  - S&P Taxonomy
  - Maps S&P use cases to Reference Architecture
- Issues:
  - Some problems are just hard
  - Need more use cases (or S&P requirements from existing)

InCadence STRATEGIC SOLUTIONS

Volume 5: Architecture White Paper Survey

- Designed to determine if there are common elements to Big Data Architecture
- Built from a survey call
  - 10 responses from Industry (8) and Academia (2)
- Was sufficient to develop a comparative view and identify key roles and functional components
  - Helped to scope top level roles in the RA
- Issues
  - Sample set was too small

InCadence STRATEGIC SOLUTIONS

Volume 6: Reference Achitecture

- Had to be Vendor Neutral and Technology Agnostic applicable to a variety of business and deployment models.
- The Goals:
  - To illustrate and understand the various Big Data components, processes, and systems, in the context of an overall Big Data conceptual model;
  - To provide a technical reference for U.S. Government departments, agencies and other consumers to understand, discuss, categorize and compare Big Data solutions; and
  - To facilitate the analysis of candidate standards for interoperability, portability, reusability, and extendibility.
- Mapped Use case categories to Reference Architecture Components and Fabrics
- Defined 7 top level and 5 sub-roles
  - Two roles presented as fabrics
- Issues
  - Hard to describe an architecture without being able to mention technologies
  - Terminology came back to bite us
  - Current architecture is not really normative
  - Too much in one diagram – mixed views

More on the RA to follow

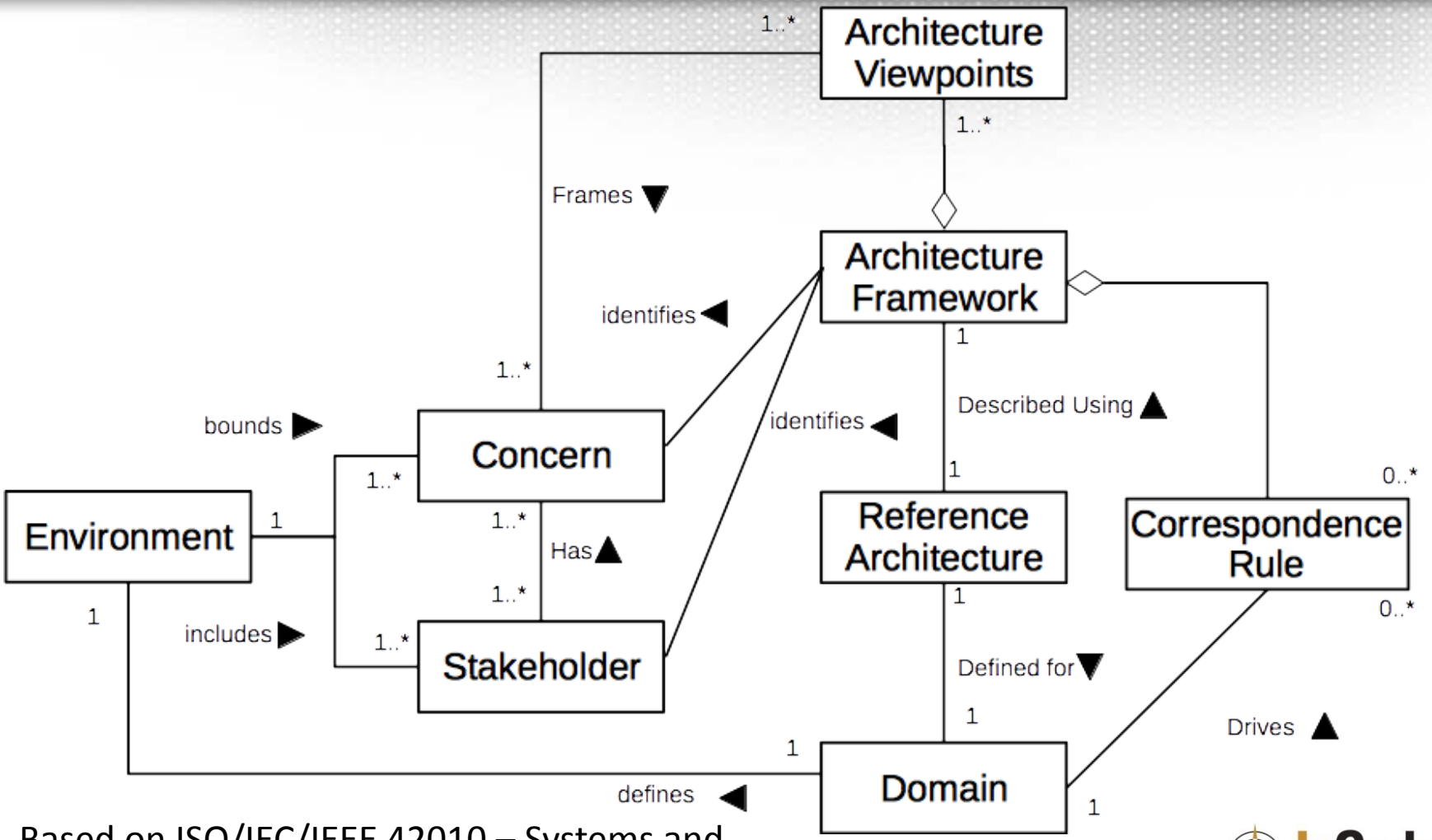
InCadence STRATEGIC SOLUTIONS

Volume 7: Standards Roadmap

- Goals:
  - Document an understanding of what standards are available or under development for Big Data
  - Perform a gap analysis and document the findings
  - Identify what possible barriers may delay or prevent adoption of Big Data
  - Document vision and recommendations
- Also designed to be a summary document
- Surveyed major SDOs and Consortium standards
  - Developed criteria for “Relevant to Big Data”
  - Mapped to Ref Arch roles as users or implementers of standard
- Issues
  - An exhaustive documentation of Big Data standards bigger than resources – Almost every standard deals with data
  - Initial direction of document was more a technology roadmap – can’t do a roadmap of technologies without mentioning technologies

InCadence STRATEGIC SOLUTIONS

Reference Architectures from a standards perspective



Based on ISO/IEC/IEEE 42010 – Systems and software engineering — Architecture description

What is the NIST BDRA?

Is

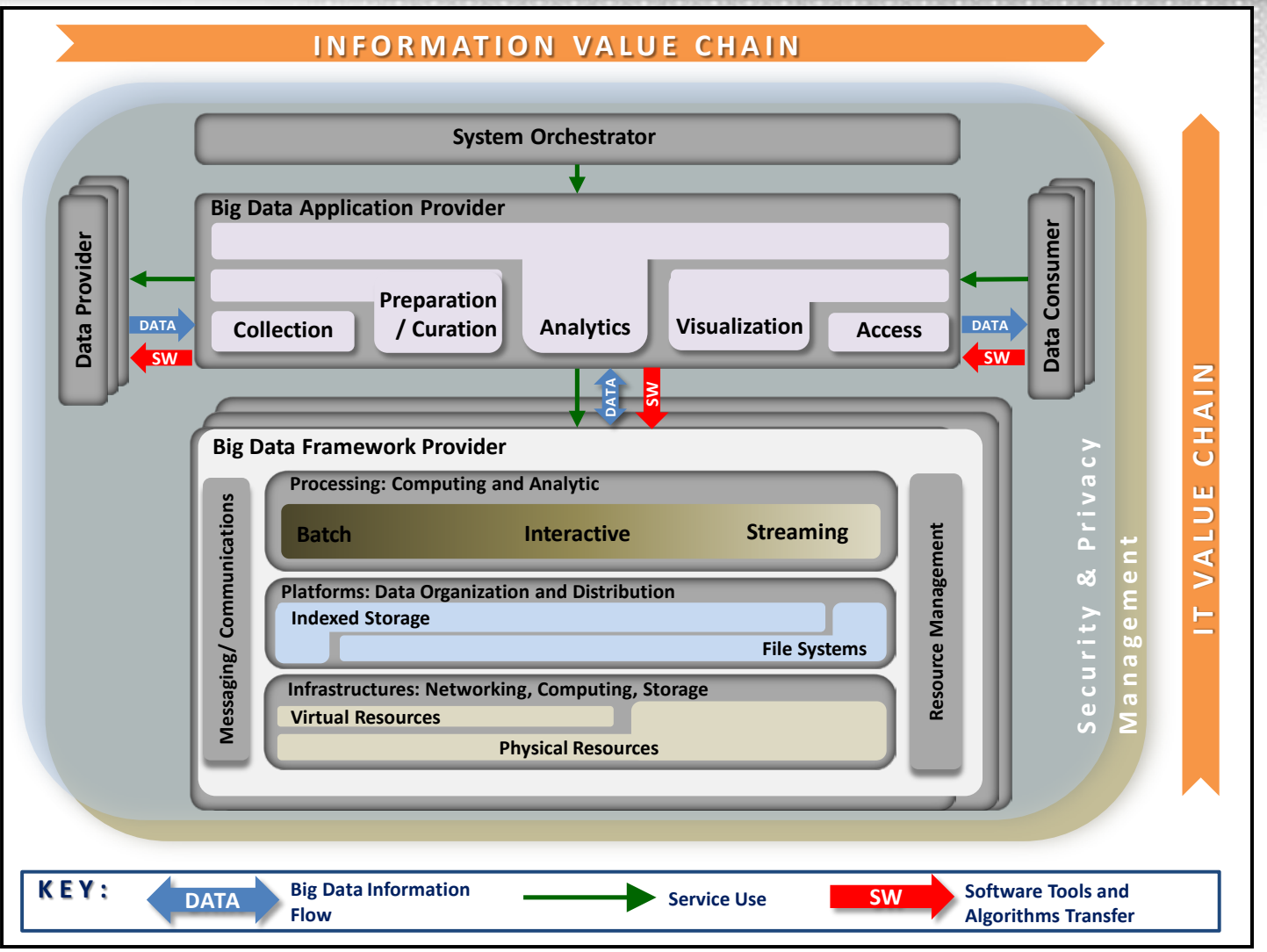
- A superset of a “traditional data” system
- A representation of a vendor-neutral and technology-agnostic system
- A functional architecture comprised of logical roles
- Applicable to a variety of business models
  - Tightly-integrated enterprise systems
  - Loosely-coupled vertical industries

Is Not

- A business architecture representing internal vs. external functional boundaries
- A deployment architecture
- A detailed IT RA of a specific system implementation

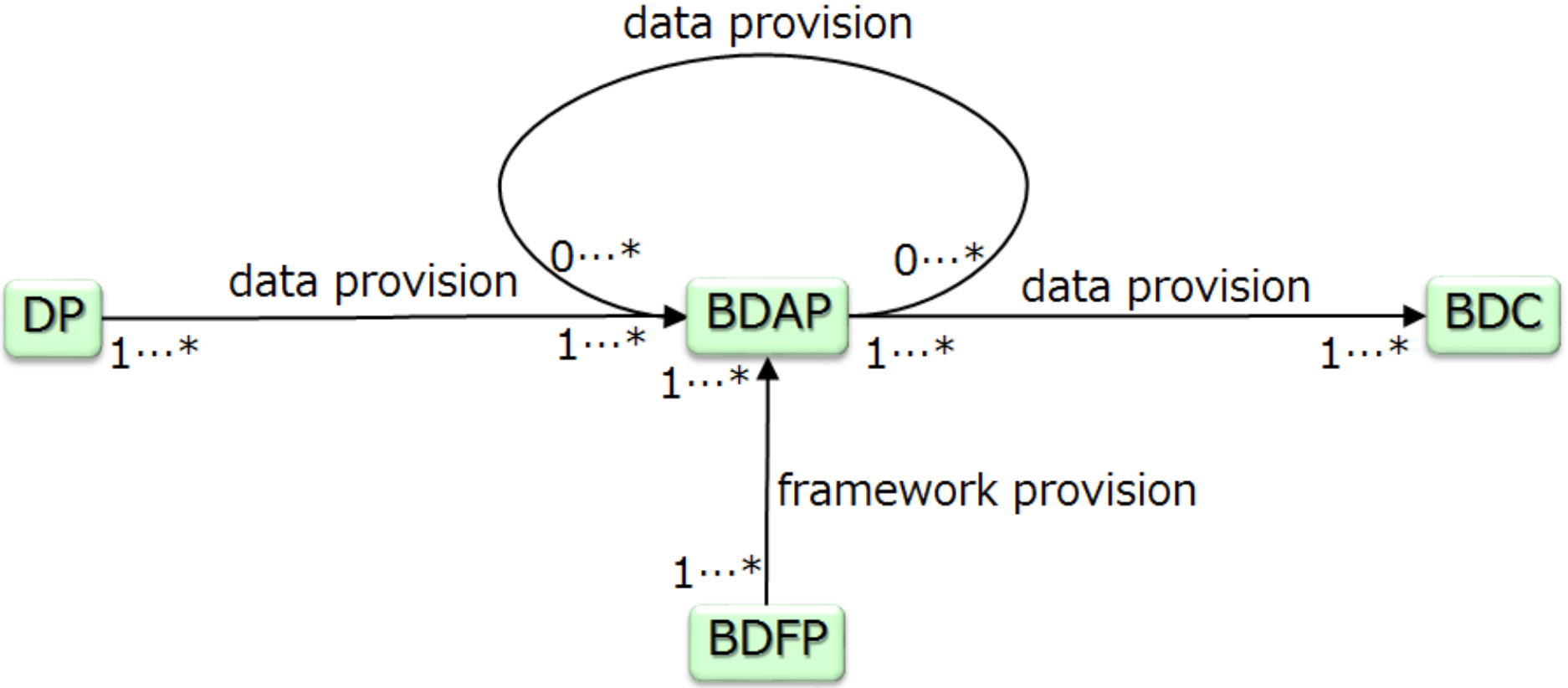
InCadence STRATEGIC SOLUTIONS

NIST Big Data Reference Architecture



InCadence STRATEGIC SOLUTIONS

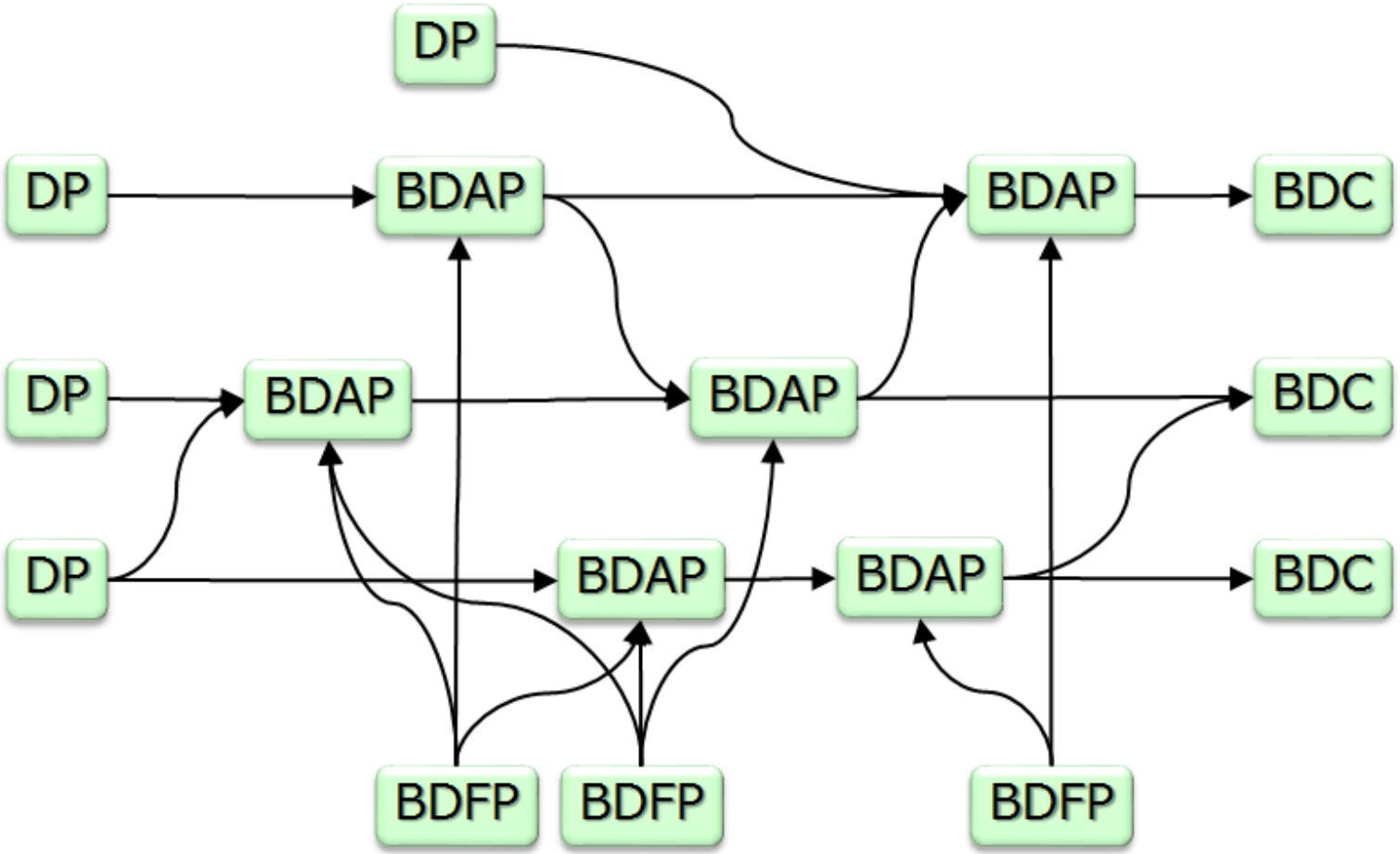
UML of the BDRA Central Roles



Courtesy of Toshihiro Suzucki – Japanese NB Representative to WG9

InCadence STRATEGIC SOLUTIONS

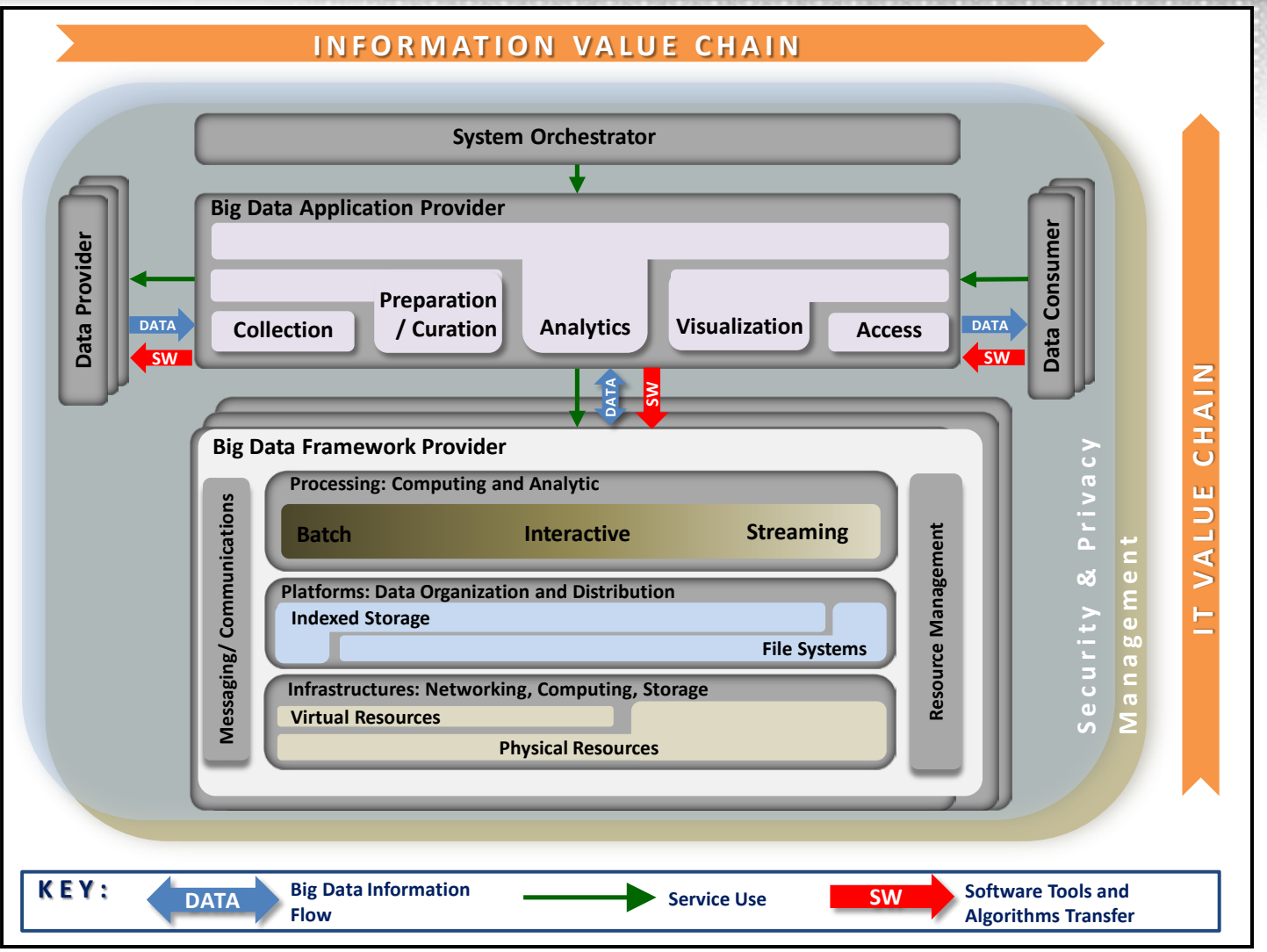
A breakout of the role relationships



Courtesy of Toshihiro Suzucki – Japanese NB Representative to WG9

InCadence STRATEGIC SOLUTIONS

NIST Big Data Reference Architecture



InCadence STRATEGIC SOLUTIONS

NIST Standards Roadmap

Standard Name/Number	Description	NBDRA Components						
		SO	DP	DC	BDAP	BDFP	S&P	M
ISO/IEC 9075-*	ISO/IEC 9075 defines SQL. The scope of SQL is the definition of data structure and the operations on data stored in that structure. ISO/IEC 9075-1, ISO/IEC 9075-2 and ISO/IEC 9075-11 encompass the minimum requirements of the language. Other parts define extensions.		I	I/U	U	I/U	U	U
ISO/IEC Technical Report (TR) 9789	Guidelines for the Organization and Representation of Data Elements for Data Interchange		I/U	I/U	I/U	I/U		
ISO/IEC 11179-*	The 11179 standard is a multipart standard for the definition and implementation of Metadata Registries. The series includes the following parts: • Part 1: Framework • Part 2: Classification • Part 3: Registry metamodel and basic attributes • Part 4: Formulation of data definitions • Part 5: Naming and identification principles • Part 6: Registration		I	I/U	I/U		U	

Implementer: A component is an implementer of a standard if it provides services based on the standard (e.g., a service that accepts Structured Query Language [SQL] commands would be an implementer of that standard) or encodes or presents data based on that standard.

User: A component is a user of a standard if it interfaces to a service via the standard or if it accepts/consumes/decodes data represented by the standard.

InCadence STRATEGIC SOLUTIONS

ISO/IEC JTC1 WG9: Terms of Reference

1. Serve as the focus of and proponent for JTC 1's Big Data standardization program.
2. Develop foundational standards for Big Data ---including reference architecture and vocabulary standards---for guiding Big Data efforts throughout JTC 1 upon which other standards can be developed.
3. Develop other Big Data standards that build on the foundational standards when relevant JTC 1 subgroups that could address these standards do not exist or are unable to develop them.
4. Identify gaps in Big Data standardization.
5. Develop and maintain liaisons with all relevant JTC 1 entities as well as with any other JTC 1 subgroup that may propose work related to Big Data in the future.
6. Identify JTC 1 (and other organization) entities that are developing standards and related material that contribute to Big Data, and where appropriate, investigate ongoing and potential new work that contributes to Big Data.
7. Engage with the community outside of JTC 1 to grow the awareness of and encourage engagement in JTC 1 Big Data standardization efforts within JTC 1, forming liaisons as is needed.

InCadence STRATEGIC SOLUTIONS

ISO/IEC 20546, Information technology -- Big Data -- Overview and Vocabulary

- Scope: This International Standard provides an overview of Big Data, along with a set of terms and definitions. It provides a terminological foundation for Big Data-related standards.
- Schedule
  - Current: 1st WD Available
  - CD: Oct 2016
  - Publication: Oct 2018

InCadence STRATEGIC SOLUTIONS

ISO/IEC 20547, Information technology -- Big Data Reference Architecture

Part	Title	Scope
1	Framework and Application Process*	This technical report describes the framework of the Big Data Reference Architecture and the process for how a user of the standard can apply it to their particular problem domain.
2	Use Cases and Derived Requirements*	This technical report decomposesa set of contributed use cases into general Big Data Reference Architecture requirements.
3	Reference Architecture	This International Standard specifies the Big Data Reference Architecture (BDRA). The Reference Architecture includes the Big Data roles, activities, and functional components and their relationships.
4	Security and Privacy Fabric	This international standard specifies the underlying Security and Privacy fabric that applies to all aspects of the BDRA including the Big Data roles, activities, and Functional components
5	Standards Roadmap*	This technical report will document Big Data relevant standards, both in existence and under development, along with priorities for future Big Data standards development based on gap analysis.

* Technical Report

InCadence STRATEGIC SOLUTIONS

Big Data Standards
Future Possibilities

- Vocabulary
  - More Formal Definitions
  - Broader more Formal Taxonomy
- Reference Architecture
  - Adopting Multiple Views
    - Activity/User
    - Functional Components
  - Pro-forma conformance process
  - Fixing Issues
    - Multiple application providers
    - Broaden/Split System Orchestrator Role
  - Tie closer to ISO/IEC/IEEE 42010 - 42010 – Systems and software engineering — Architecture description
  - Map Use Cases
  - Document Interface Patterns

InCadence STRATEGIC SOLUTIONS

Conclusion

- The BDRA can provide a framework to consistently describe Big Data system
  - Roles
  - Activities
  - Components
- Standards Roadmap can help guide selection of standards requirements
- We need help to make the process and the products better:
  - NIST BDPWG
  - INCITS TC Big Data
  - ISO/IEC JTC1 WG9

InCadence STRATEGIC SOLUTIONS