Towards A General Reference Architecture for BIG DATA

Creating A Formal Specification

By Gary Mazzaferro, AlloyCloud Nov.2011

Updated: 07.2013

BIG DATA General Reference Architecture:

NOTICE:

This Document Contains Selected Slides Relevant To NIST's BIG Data Initiative

The Slides In This Document Were Originated By AlloyCloud
And Disclosed To U.S. Goverment Departments And Agencies Under IPR

Only Information Presented This Document May Be Considered
In Public Domain

Confidentiality Agreements Remain In Force

BIG DATA General Reference Architecture:
Contents

BIG DATA General Reference Architecture:
Big Data Simple Definition

- Big Data Is A Shift In the Way We Consume, Process And Apply Information To Create Intelligence.
- Approach:
  Take Advantage Many Data Sources To Expose Hidden Knowledge Lost In Tradition Data Processing
- How:
  Employing Social Media, Text Processing, Natural Language Processing.. Flexible/Dynamic Database Schemes
- While:
  Often Bypassing Tradition Tools, Policy And Processes Accelerating Results

BIG DATA General Reference Architecture:
Why A Hot Topic ?

- Growing Interest And Attention Created By The Demand For New Kinds Of Actionable Intelligence.
- Being Enabled By The Intersect Of Modern Communications And Compute Capabilities Reaching Into World Of Pervasive, Ubiquitous Data
- Made Real By A New Technical Generation, Unconstrained By Tradition, Rediscovering Approaches Deemed Unfeasible And Too Costly in the '80s.
- High Velocity Value Delivery -- Very Scalable

BIG DATA General Reference Architecture:
Industry Challenges

- New Approach Where Implications Are Not Well Understood
- Technologies Are Emerging With An Unproven Track Record
- How To Maximize Big Data Value Delivery
- Customers Concern On Procurement
  - Which Products Are Best Fits ?
  - How To Integrate ?
  - Skill Sets To Operate And Maintain Big Data Systems

BIG DATA Cloud Enterprise Resource Framework:
BIG DATA Unique Capabilities

- Varied Ingest Data Schemes
- Processing Social and Other Media Sources
- Request Semi-Structured Data In Queries
- Integrated Autonomic Partitioning & Workload Planning (some technologies)
- Independent Cluster Style Workload Balancing (some technologies)
- Independent Data Distribution Across Clusters Nodes (some technologies)
- Open Source Flexibility – "Just Code" A New Feature
- New Query Languages Emerging
- RESTful, Web Based Access Protocols (some technologies)

BIG DATA Cloud Enterprise Resource Framework:
BIG DATA Myths

- BIG DATA Is A New Idea
- BIG DATA Automatically Discovers New Knowledge
- BIG DATA Is A Standard
- BIG DATA Is Cloud Computing
- Map Reduce Is BIG DATA
- BIG DATA Provides Multi-Tenant Security
- BIG DATA Generates Standard Reports
- BIG DATA Is Low Cost
- BIG DATA Is Real Time
- BIG DATA Processing Over The Public Internet

BIG DATA Cloud Enterprise Resource Framework:
BIG DATA Myths Dispelled

- BIG Data Is A New Idea FALSE
  - In the 1980s It Used to be Called "Distributed Database Management System" (DDBMS)
  - The Techniques Are The Same: Query Load Balancing, Range partitioning, Composite partitioning, Vertical partitioning, Horizontal partitioning (sharding)
- BIG DATA Automatically Discovers New Knowledge FALSE
  - BIG DATA does not auto-magically find new information
  - A data scientist must analyze each data source and programmers must the code for data processing
- BIG DATA Is A Standard FALSE
  - Today, There are NO International Standards for BIG DATA
  - Vendors Claim Apache Hadoop Is a "Defacto Standard". Unfortunately It Only Works for "Hadoop BIG DATA"
  - BIG DATA May Leverage Other Standards. However, There Are NO Minimum Compliance Profiles for BIG DATA
- BIG DATA Is Cloud Computing FALSE
  - Cloud Computing Is a WAY of Procuring Compute Resources
  - BIG DATA Can Be Deployed On Cloud Infrastructures OR Clusters, Mainframes Traditional Compute Infrastructures
- Map Reduce Is BIG DATA FALSE
  - Map Reduce Is ***Only One Of Many*** Cluster Computing, Load Balancing Techniques Used by Some BIG DATA Technologies
  - Map Reduce is NOT a Requirement for BIG DATA
- BIG DATA Provides Multi-Tenant Security FALSE
  - Today, Multi-tenancy Is Not Considered Part Of BIG DATA
- BIG DATA Generates Standard Reports FALSE
  - BIG DATA Technologies Have NO Standards Reports
  - All Reports Must Be Created By Data Scientists and Programmers
- BIG DATA Is Low Cost FALSE
  - Text and Natural Language Processing Can Consumes a High Number of CPU Cycles Driving Up Costs
  - Infrastructures Require Extreme Network Bandwidth Driving Up Costs
  - Text and Natural Language Processing Intermediate Results Is Usually Kept In High Performance Storage Driving Up Costs
  - Technologies Are Extremely Complex and Difficult to Operate Without Procuring Costly Support Contracts
- BIG DATA Is Real Time FALSE (mostly)
  - Real-Time Is Subjective, If Data Processing Meets Delivery Requirements, It Is Real-Time
  - Text and Natural Language Processing Can Take a High Number of CPU Cycles With Unpredictable Completion Times
- BIG DATA The Public Internet FALSE (mostly)
  - Don't Expect Petabytes of Data Processing to Occur Overnight Using the Public Internet and Low Cost Cloud Computing
  - 1TB of data will take 500-1000hrs to read using a 100mbs network connection. That is 3-6months not including temporary results storage.
  - Many BIG DATA Technologies Cannot Operate In A WAN environment.

BIG DATA Cloud Enterprise Resource Framework:
BIG DATA Commercial Enterprise Key Capabilities (should have wish list)

- Information Interoperability – Any Information From Anywhere
- Identify Same Data Across Different Sources and Time-Shifted From Same Source
- Autonomous Self-Healing Storage/Compute Infrastructure
- Autonomous, Policy Based, Comprehensive Workload Management
  - Signal & Natural Lang. Proc, Work Locations, Users, Jobs, Completion Dates
- Autonomous System Optimization – App Profiles, Data/Data Processing/Network Performance Tiers
- Standard Capabilities Catalog
- Interoperability Across Vendors Products
- Common, System Wide Event Reporting & Logging
- Application Optimize Through Selecting Best of Breed Technologies
- Reference Architectures – Guides Planning, Design and Deployments

BIG DATA Cloud Enterprise Resource Framework:
BIG DATA Defense/Intelligence Additional Key Capabilities

- Generalized Capabilities - not application or program specific)
  - Data Anomaly Detection (ADAMS) (Tampering, Errors, Inconsistencies, Age/Currency)
  - Anomaly Tolerant Query (non-Stochastic, non-Causal Query)
  - Information/Data Confidence Maturity Models
  - Autonomous Security Threat Response and Reporting
  - Multi-Lateral, Multi-Level, Authentication, Authorization, Confidentiality Information Security -Supports Redaction (Dynamic ABAC On Steroids)
  - Real-Time Information Redaction -e.g. Video, Imaging, Audio, Text, File, DB Records, Documents, Paragraphs Sentences, Phases, Words, Personal Information, Other Sensitive Information, Meta-Data
  - High Granularity Data Management – Search, Resilience, Provenance, Geo-location, Replication, Confidentiality, Maturity Models, Life Cycle – Near-line, Offline, Archival, Destruction
  - Processing Using Encrypted Code At Data Site
  - Processing Encrypted Data
  - Operation Over Low Bandwidth, Intermittent, Low Integrity Communications Networks
  - Access to Other Resource Sources - Scientific Grid, OOI, Web Compute Resources (Other Depts Agencies, NGOs, Foreign Govt Agencies, Coalition Partners)
    - e.g. FAA, DOE, NARA, NIH, FEMA, DOI, Foreign Govt. Agencies, Red Cross, Police, Firefighting, Local Volunteers, Municipal Transit, Private Doctors, Pharmacies, Hospitals, Ambulance Services, Oil/Fuel Distribution
  - Alignment With Net-Centric Approaches
  - DoDAF styled BIG DATA Reference Architectures

BIG DATA General Reference Architecture:
BIG DATA Commercial Enterprise Key Gaps & Short Comings

- Resource Planning, Deployment, Optimization and Costs
  - Semi-Structured Data Processing Unpredictable Completion Times Makes Scaling, Resource And Budget Planning Difficult
  - BIG DATA Proprietary QLs – Competency/Talent Gap, Rewrite Legacy SQL Reports/Queries, Rewrite Data Warehouse Queries
  - No Best Practices Regarding Applications, Architectures, Operations and Deployments
  - Disconnected Management, Administration and Deployment Tools from Mainstream Drives Up OpEx and Reduces Agility
  - NO Alignment and Leverage with Cloud Data Mgt/Access Standards Without Significant Custom Development
  - Each Unique Data Source Requires Custom Development, Costly Data Scientists Required
  - NO Trade-Off Model for "On the Fly vs. Stored" Denormalized vs Normalized Data
  - NO Integrated Chargeback Tracking/Reporting/Billing for Resource Consumption e.g. Service Levels, Tiers, In Plan, Out Plan
  - BIG DATA Can Be Too BIG To Moved Via Networks From Place of Residence, May Require "Secure Agent Based" Data Processing
- Quality and Data Integrity
  - Emerging Technologies --- NO Quality of Record
  - Poor Leverage/Integration with Existing Storage Infrastructure Management, Data Management and Disaster Recovery
  - BIG DATA Tech. NOT HARDENED, Open Source Funding, Sub-Optimal Reliability ("Kindness of Strangers" Quality Model)
  - NO System Wide Diagnostics i.e. Execution Logging and Traceability, Logging Proprietary per Technology
- Management, Administration and Interoperability
  - BIG DATA Tech. Load Balancing Not Integrated to Cloud/GRID/Cluster Workload Management Tools
  - No Consistent/Common Management and Common Monitoring and SLAs NON-Existent Across BIG DATA Technologies
- Security
  - Authorization Privileges and Enforcement NOT Consistent Across BIG DATA Technologies
  - NO Integrated Third-Party Service/Partner Credential Management

BIG DATA Cloud Enterprise Resource Framework:
BIG DATA Defense/Intel. Additional Key Gaps & Short Comings

- Resource Planning, Deployment, Optimization and Costs (generalized- not application or program specific)
  - Proprietary APIs and Mgt Tool Make Optimizing Applications and Technology Adoption Cost Prohibitive
  - NO Reference Architectures to Guide Deployments e.g. Strategic, Applications, Cloud, Partner Interoperability
  - Each Unique Data Source Requires Custom Development, Costly Data Scientists Required
  - NO Trade-Off Model for "On the Fly vs. Stored" Denormalized Data
  - NO Time Deadline Based Resource Provisioning, Acquisition and Workload Management
  - NO Workflow Synchronization to External Systems and No Control of External Data Processing Without Custom Development
  - NO Knowledge/Information/Data Virtualization and Interoperability Standards: New data Types Require Custom Development
  - NO Comm. Channel to Data Type Awareness and Over Low Bandwidth, Intermittent, Low Integrity Communications Networks
- Quality and Data Integrity (generalized- not application or program specific)
  - BIG DATA TECHNOLOGIES ARE NOT DESIGNED FOR LIFE-CRITICAL APPLICATIONS
  - BIG DATA Intolerant Intermittent Data Availability and Anomalous Data and Data Processing
  - NO Integrity Management –ie confidence models, currency models, monitoring and data validation, "End to End" Data Integrity Enforcement, Config. Mgt
  - NO System Resiliency Repair, Recovery and Validation Tooling
- Management, Administration and Interoperability (generalized- not application or program specific)
  - Query and Search, Catalogs, Languages Inconsistent and DO NOT Interoperate Across BIG DATA Technologies
  - Query Results DO NOT Interoperate Across BIG DATA Technologies Without Custom Development
  - No Interoperation with Standards: Cloud, Data Management, Storage Management, Deployment Configuration,
  - No Standards for Capability, Service and Data Catalogs: Joint, Packages, Coalition Contribution
- Security (generalized- not application or program specific)
  - NO Integration with Third-Party AA/Confidentiality Systems e.g. User, Rank, Clearance, Partner, Storage, Partner/Vendor Data Services, Multi-Tenant
  - No Granular Confidentiality On Data, Multi-Tenant Isolation/Secure Separation
  - NO Threat/Data Tampering Detection, Std. Reporting and Response
  - NO Processing Encrypted Data and Encrypted Queries
  - NO Granular Redaction for Raw Data, Queries and Reports ie Video, Imaging, PII, Scans, Documents, Paragraphs, Text, Audio
  - NO Dynamic Authorization e.g. Geo-Location, Access Device, Environment Risk

BIG DATA General Reference Architecture: Reference Architecture

BIG DATA General Reference Architecture: Reference Architecture Purpose

- Structured Approach To Document And Communicate Aspects of a "System"
- Focuses On Solutions
- Identifies Key Areas Of Interest (KAI)
- Elaborate Relationships, Touch Points And Interactions
- Provides A Framework For Technical Solutions
- Employs Views To Impart Specific Concepts In A Way Readers Are Accustomed

BIG DATA General Reference Architecture: Community Reference Architecture Differences

- Founded In Industry Ecosystems
- Community Participation Approach With Open, Transparent, Consensus Base Activities
- Incorporates A Broad Stakeholder Consensus Base
- Provides A Stereotypical Solution (Guidance) By Aggregating Known Solutions and Anticipated Capabilities
- Product An Technology Agnostic

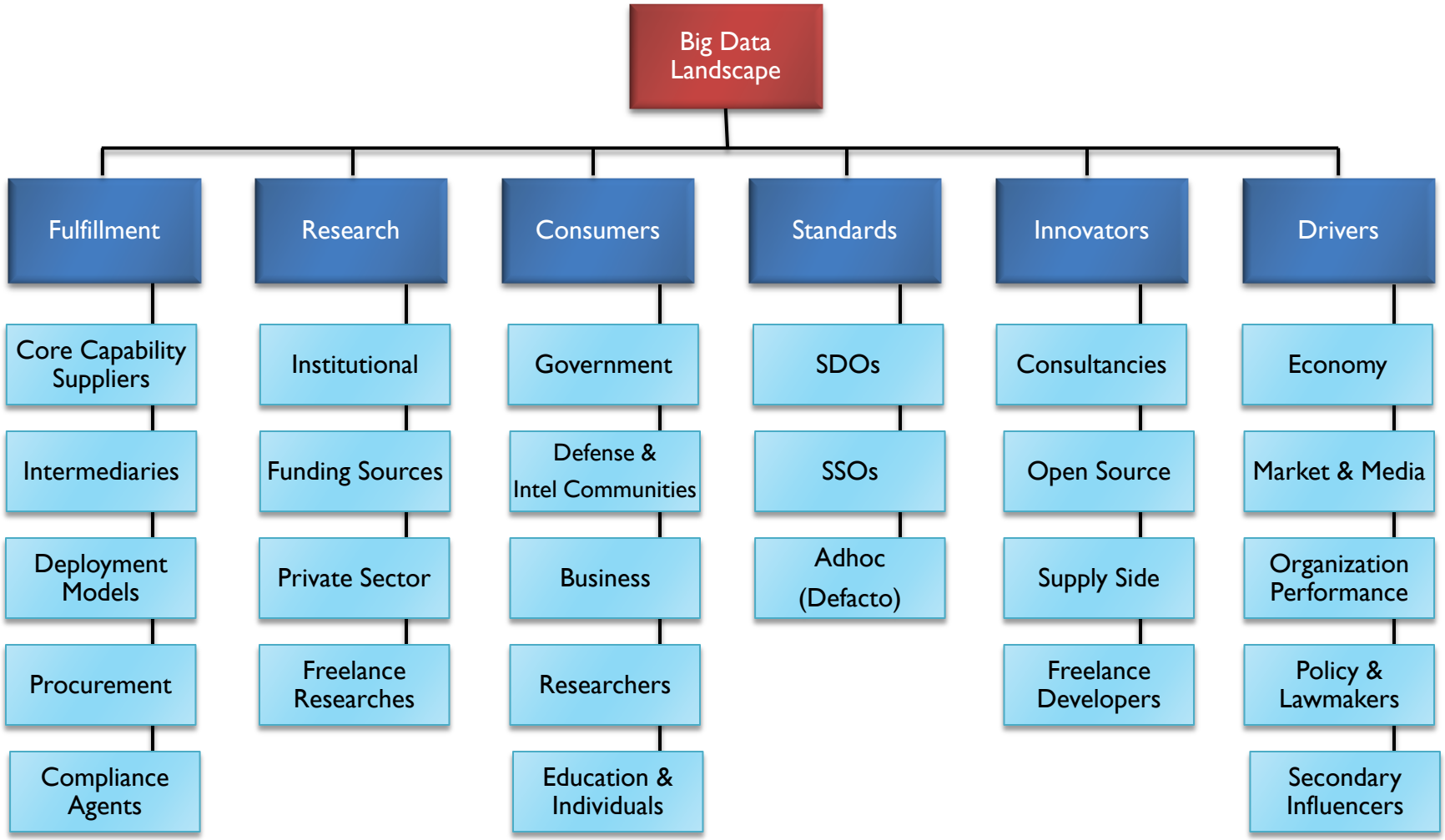
BIG DATA General Reference Architecture: NIST Cloud Initiative Lessons Learned

- Not One Reference Architecture View Meets All Stakeholders' Needs
- New Views Causes Unsure Stakeholder Interpretation And Requires Investment In Stakeholder Education
- Stakeholders Requested Actionable Technical RAs
- RAs and Language Sensitivity Promoting Solution Inclusion And Avoid Exclusionary Language

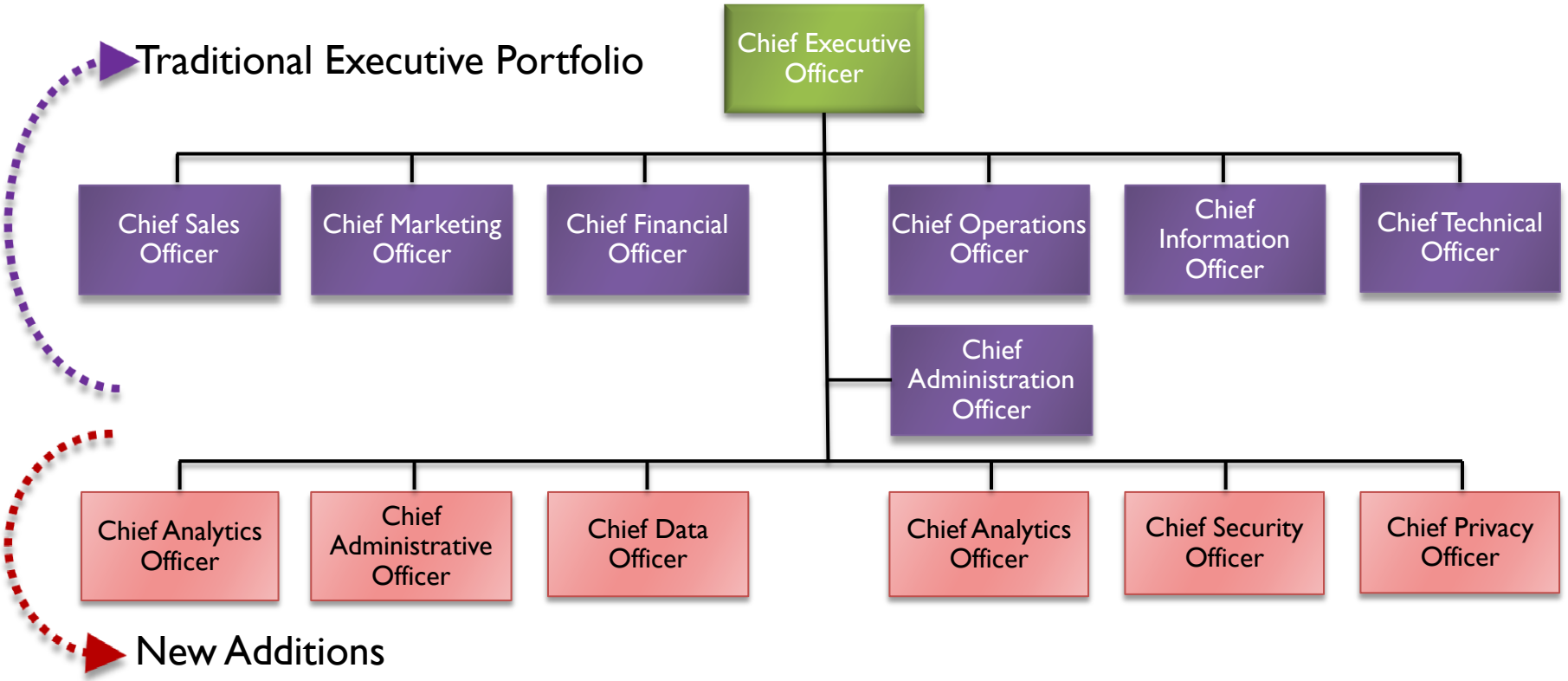
BIG DATA General Reference Architecture: General Reference Architecture Debate

- Debate Surrounding A Standard "General Reference Architectures"
  - Processes And Methods to Create
  - Stakeholders Perspectives
  - Scope – Which Concerns And Key Priorities
  - Views – Which Aspect(s) To Include
  - Granularity Depth of Information
  - Expressions – Diagrams, Document Flows And Formats
- Research
  - Generic Enterprise Reference Architecture (GERA)
  - Generic Enterprise Engineering Methodology (GEEM)
  - Generic Enterprise Modeling Tools and Languages (GEMT&L)
  - Generic Enterprise Modules (GMs)
- Industry Response --Guidance's
  - DoD "Reference Architecture Description"
  - Net Centric
  - SOA
  - FEA

BIG DATA General Reference Architecture: Top Level Eco-System Stakeholders



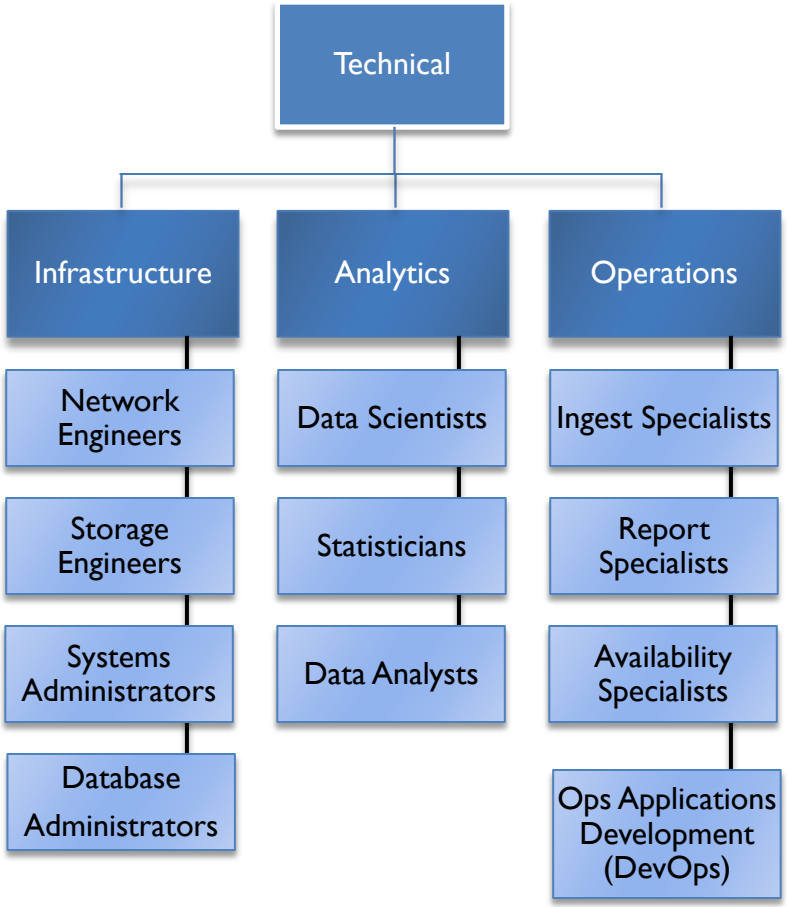
BIG DATA General Reference Architecture: Top Level Organizational Stakeholders



Response to New Opportunities and Concerns In Shifting Business and Social Landscape

Note: Procurement May Fall Under Either CFO, COO, CAO, CEO Responsibilities

BIG DATA General Reference Architecture: Top Level "Special Interest" Stakeholders



BIG DATA General Reference Architecture: Top Level Stakeholder Summary

- As we can see, the Big Data Landscape Creates An Ecosystem Rich With Diversity
- Today, There's No Clear Understanding Of How Big Data Will Unfold Into Interested Communities
- We Can Anticipate The Emergence Of Communities With Differing Needs and Priorities Surrounding "Big Data"
- We Can Expect Big Data's Evolution And Adoption Will Occur Concurrently At Varying Velocities Within And Across Communities

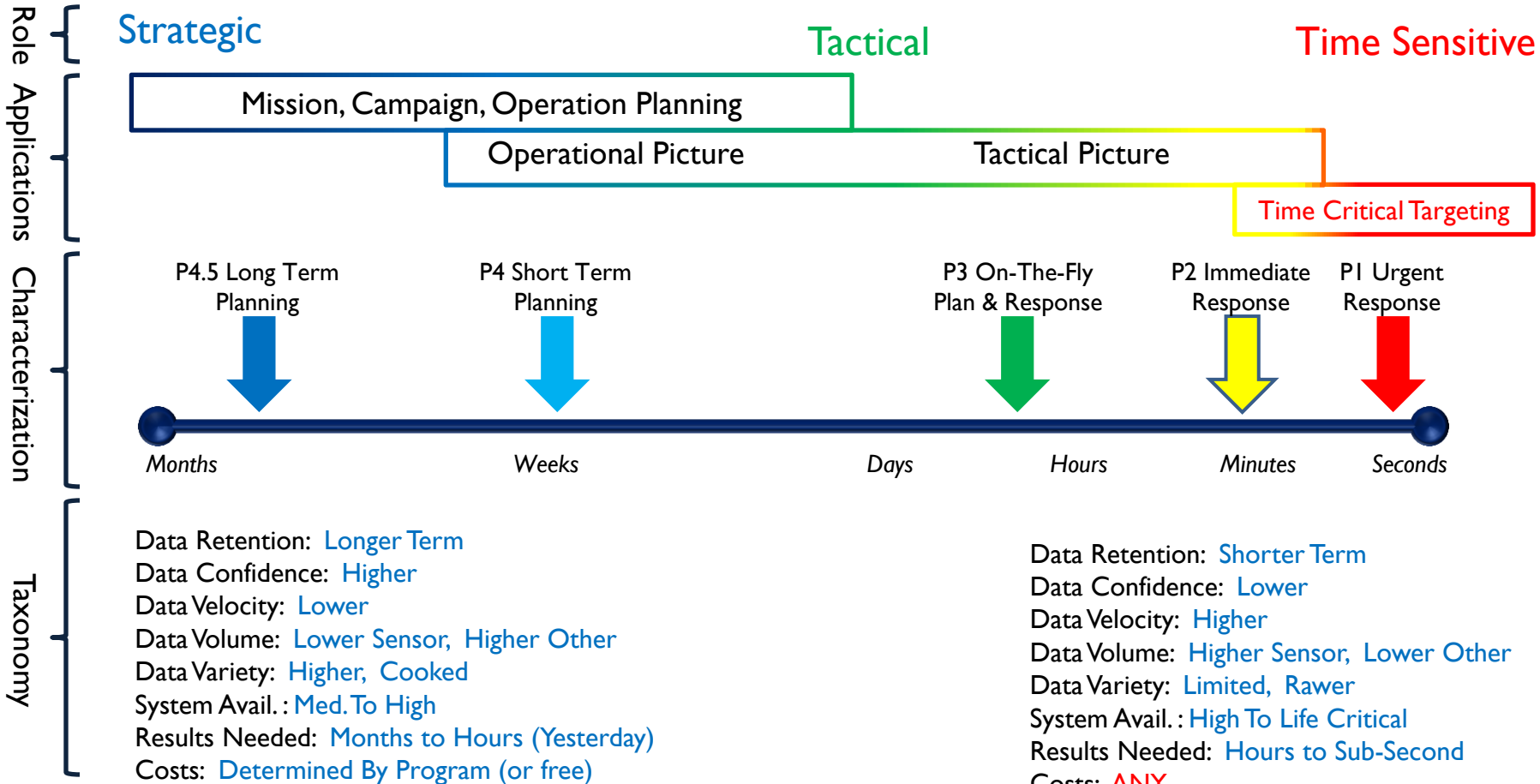
BIG DATA General Reference Architecture: Reference Architecture

BIG DATA General Reference Architecture: General Reference Architecture Views

- Eco-System
  - Aligns Market Drivers With Solutions And Participants
- Capability
  - Identifies and Aligns System Abilities
  - Facilitate Alignment To Requirements
- Technical
  - Identifies and Aligns Technical Areas
  - Defines Areas of Technical Responsibilities
  - Defines Interface Surfaces
    - Technology Agnostic
    - Data Processing Order Agnostic
- Resource Flows
  - Definition of operational concepts
  - Applying a local context to a capability
  - Allocation of activities to resources
- Deployment
  - Identifies Approaches And Options Surrounding Solution Topology
- Security
  - Aligns Security Approaches And Features With Other RA Models
- May Consider Other Reference Types and Topic Areas
  - RA of Adopted RAs
  - Processes
  - Life Cycles

BIG DATA General Reference Architecture:
Application Profile Landscape

BIG DATA Applications Have Widely Differing Operating Needs



NOTE: Anticipated Application Characterizations *(Area for Study i.e. Capabilities Catalog/ Taxonomy Spec.)*

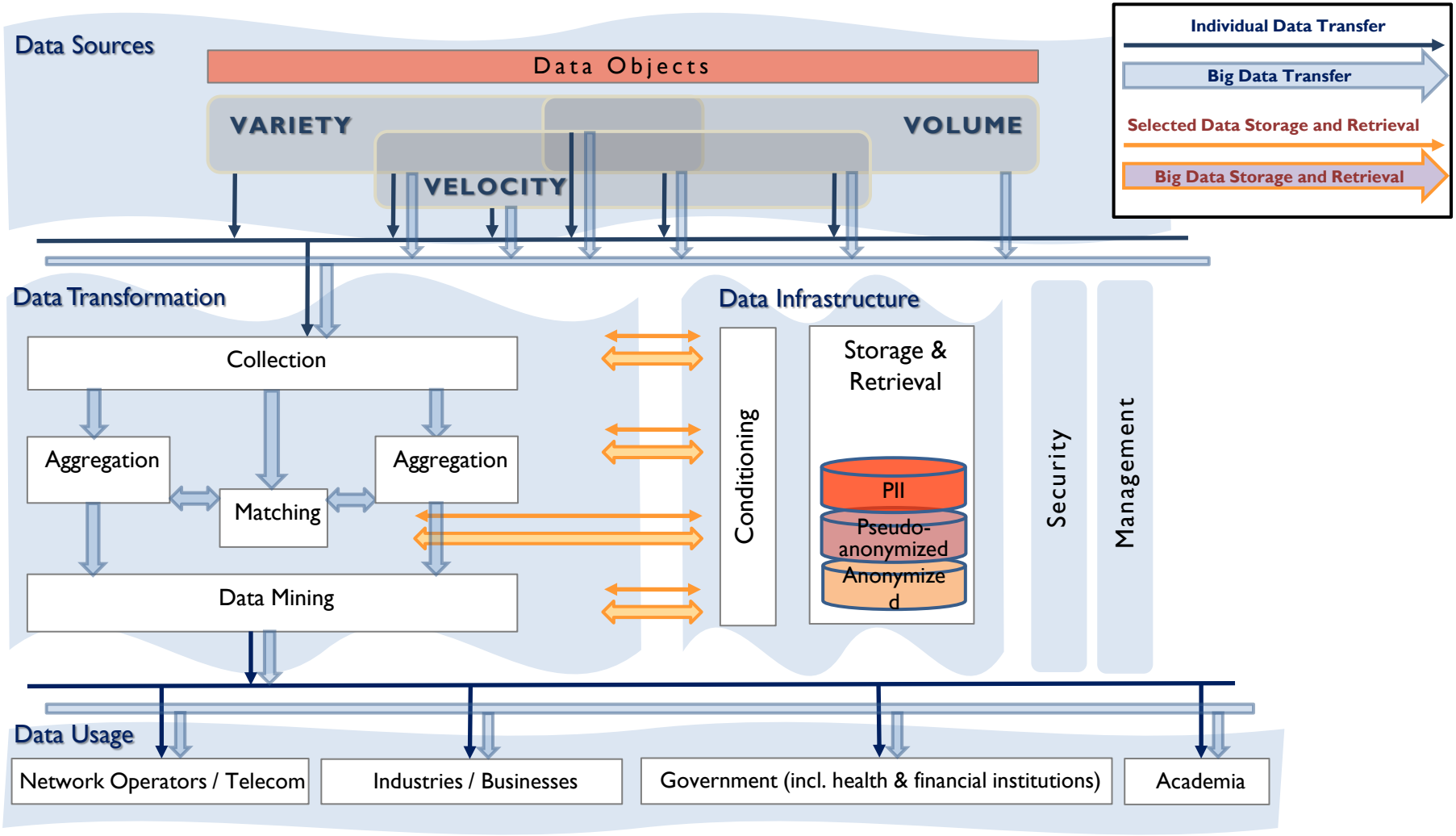
BIG DATA General Reference Architecture: Comprehensive Capabilities Taxonomy

- Transforms "Other" Capabilities Formats To A Common Reference Architecture Consumable
- General Systems Capabilities
  - Account Management And Monitoring
  - User Administration And Monitoring
  - Security
  - Federation (Models) Management And Monitoring
  - Configuration (Models) Management And Monitoring
  - Deployment (Models) Management And Monitoring
  - Availability – Metrics And Qualitative Levels (Experimental, Commercial, Mission Critical, Life Critical)
  - Procurement Compliance Management And Monitoring?
  - Maintenance & Diagnostics Management And Monitoring
  - License Management And Monitoring
  - Data Management And Monitoring
  - Supported Ingest Formats
  - Supported Output Formats
  - Supported Devices
  - Supported Interfaces
  - RA and Standards Compliance
  - Performance (Models) Management, Monitoring, Metrics And Qualitative Levels
  - User Support Capabilities- Education, Help Management And Monitoring
  - Vendor Support Capabilities - Maintenance Management And Monitoring
- System Specific Capabilities
  - Data Characterizations (Dynamics, Types of Change, Rate Of Change, Confidence, Quality, Demand)
  - Workload Management And Monitoring
  - Infrastructure Management And Monitoring (Compute Management, Storage Management, Network Management)

Nearly 500 Detailed Capabilities/Functions Defined
About 25% - 30% Complete

Note: Some Capabilities Are Functionally Cross-Cutting

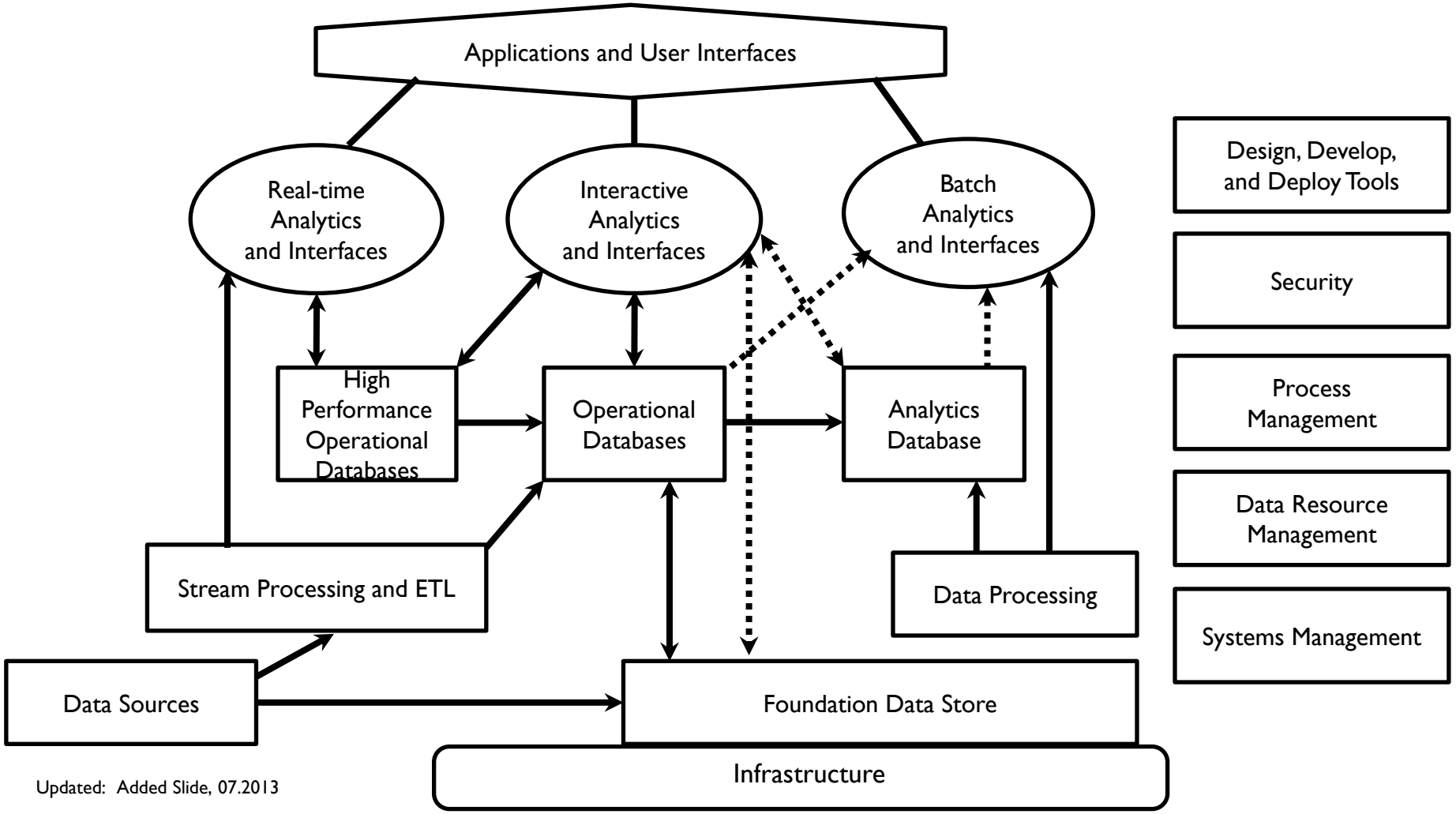
BIG DATA General Reference Architecture:
Ecosystem Viewpoint



Attribution: Orit Levin, Microsoft 07.13.2013

Updated: Added Slide, 07.2013

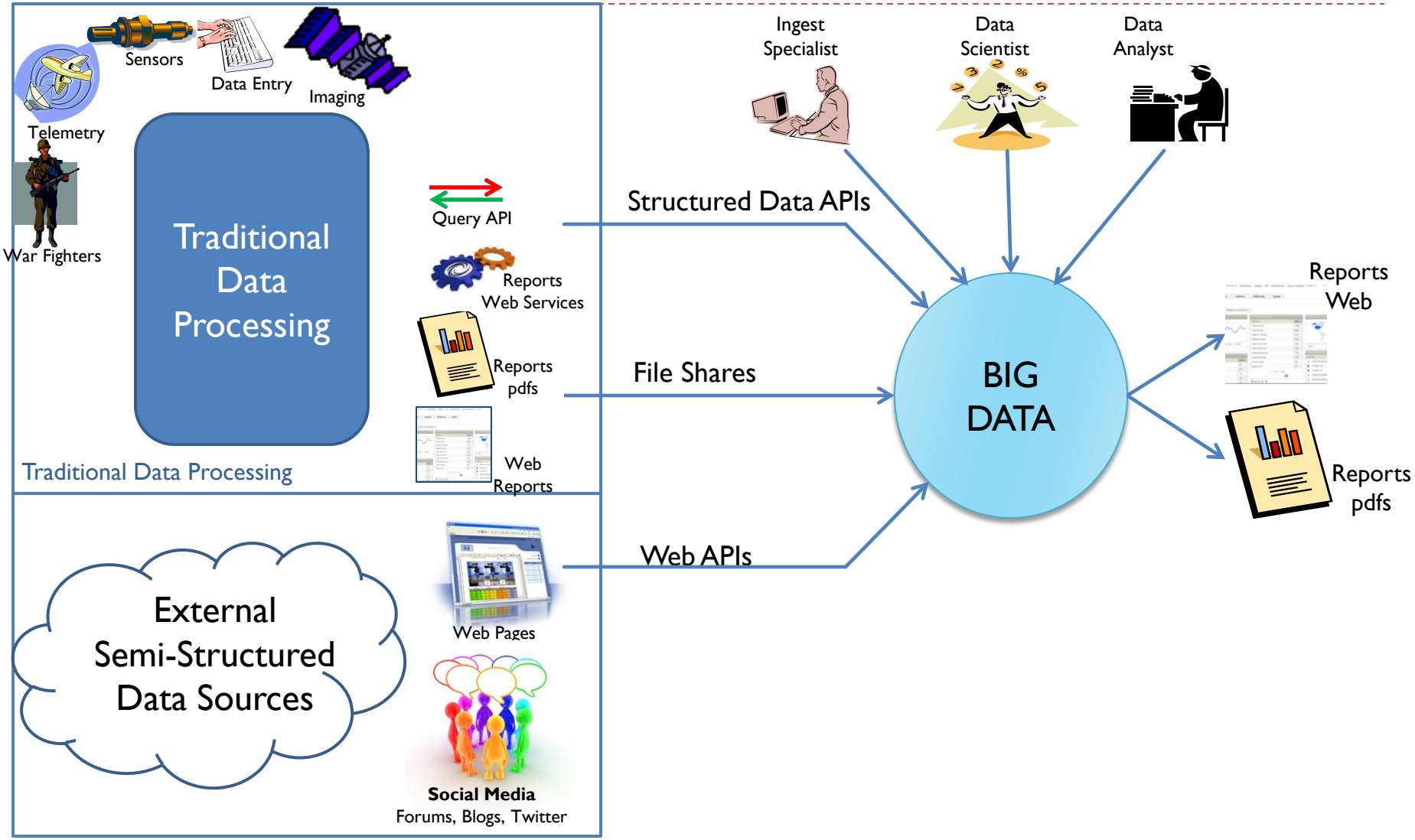
BIG DATA General Reference Architecture:
Capabilities Viewpoint



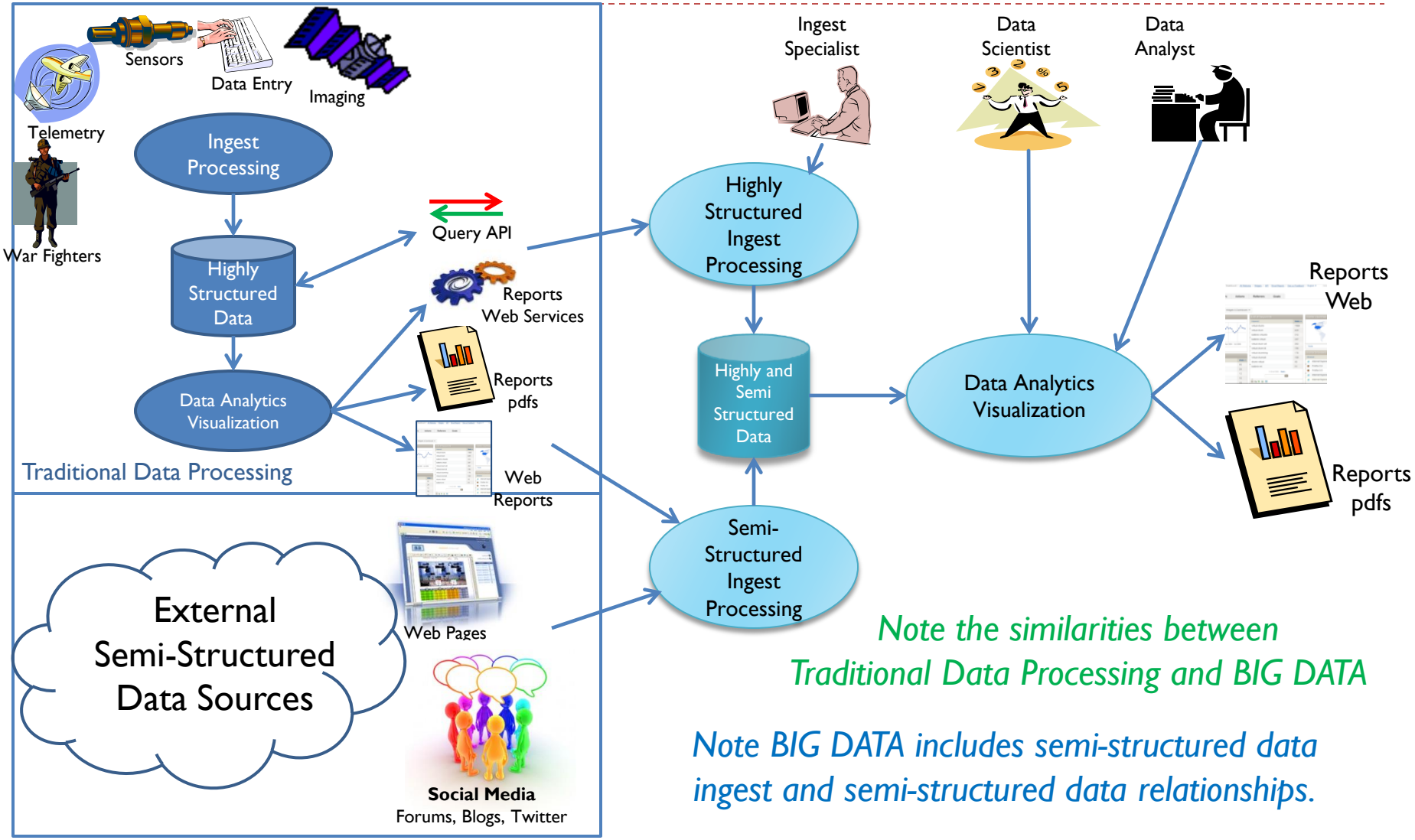
Updated: Added Slide, 07.2013

Attribution: Robert Marcus, 07.19.2013

BIG DATA General Reference Architecture:
BIG DATA High Level Operational Concepts (OV-1)



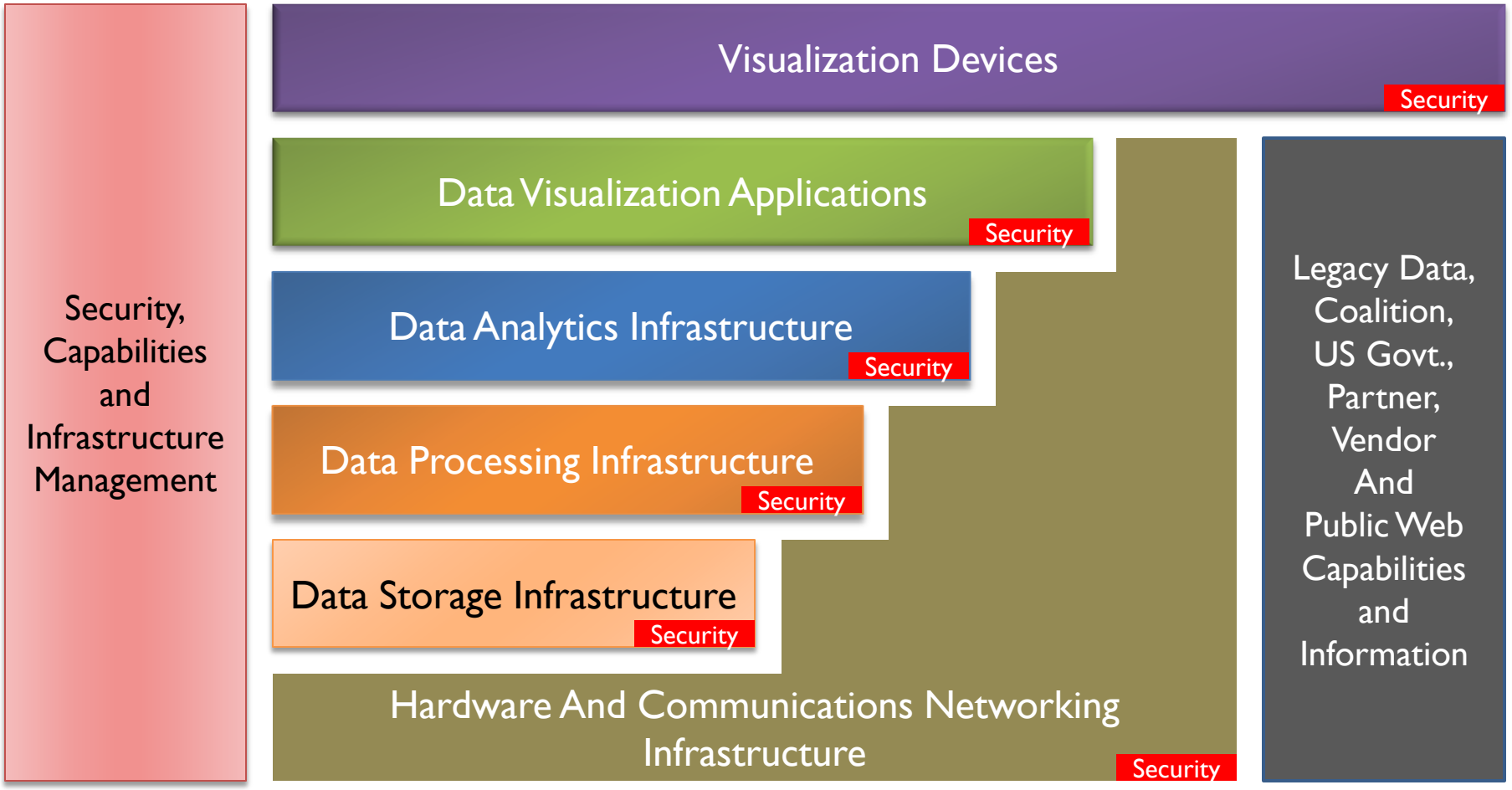
BIG DATA General Reference Architecture:
BIG DATA High Level Operational Resource Flow (OV-2)



Note the similarities between Traditional Data Processing and BIG DATA

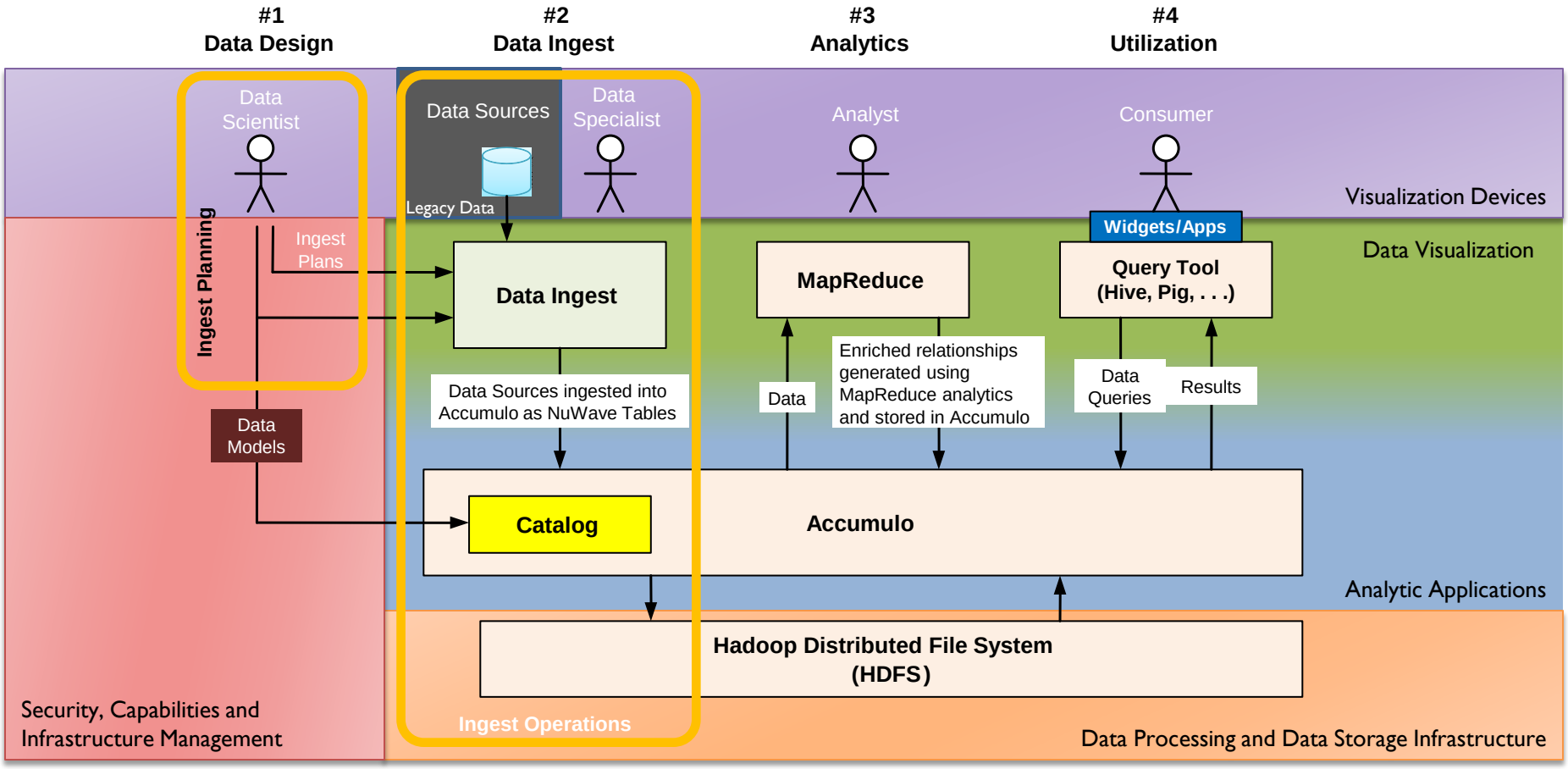
Note BIG DATA includes semi-structured data ingest and semi-structured data relationships.

BIG DATA General Reference Architecture:
Reference Architecture Technical Viewpoint



BIG DATA Common Reference Architecture:

e.g. Reference Architecture Mapped to Accumulo/Hadoop

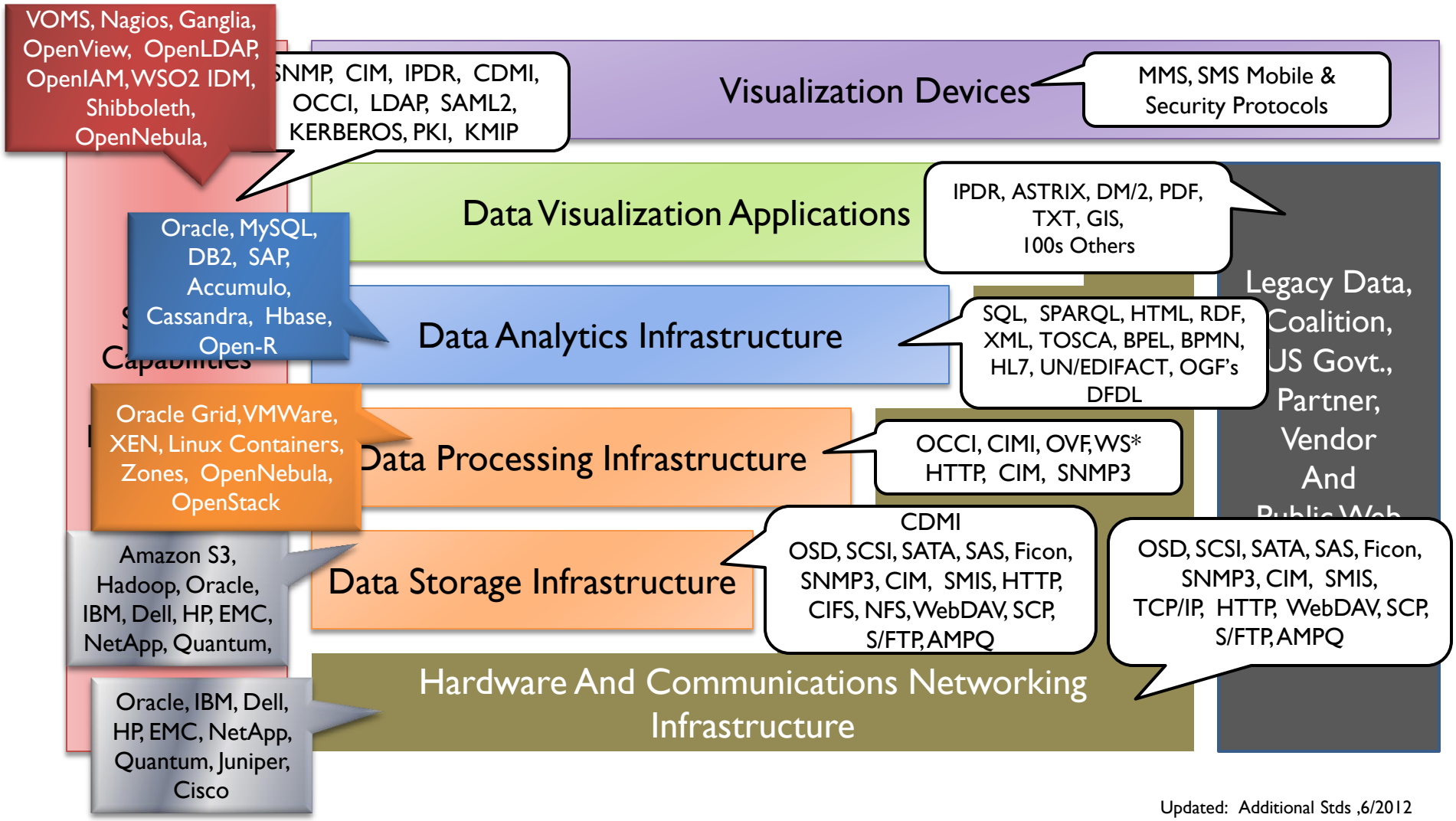


Updated: Added Slide,11/2012

Accumulo/Hadoop Attribution: "Big Data from a DoD Perspective 0.2"

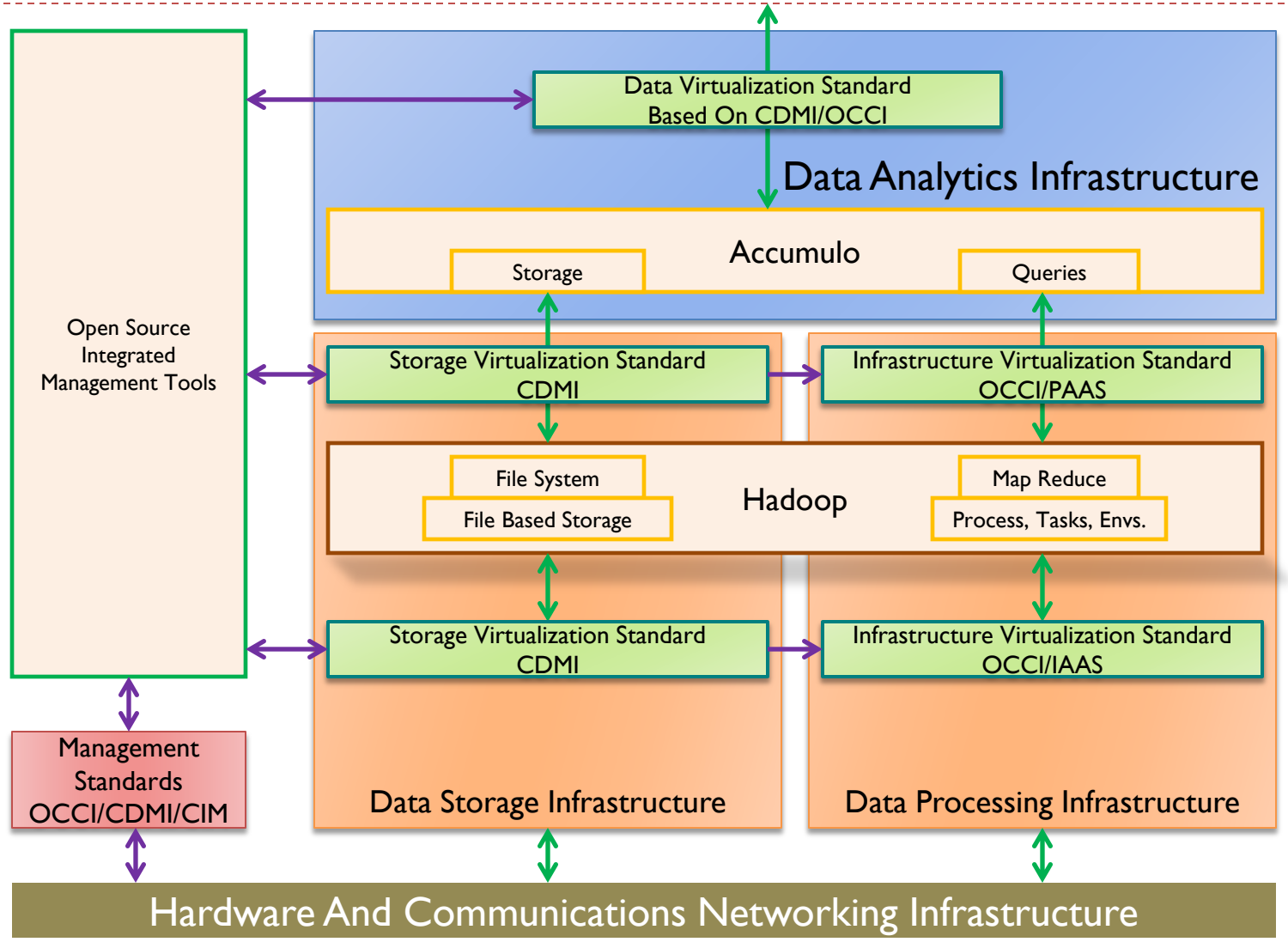
BIG DATA Common Reference Architecture:

Example: RA Technical w/ Applicable Standards And Apps



Updated: Additional Stds ,6/2012

BIG DATA Common Reference Architecture:
Example: Hadoop/Accummulo Using Applicable Standards



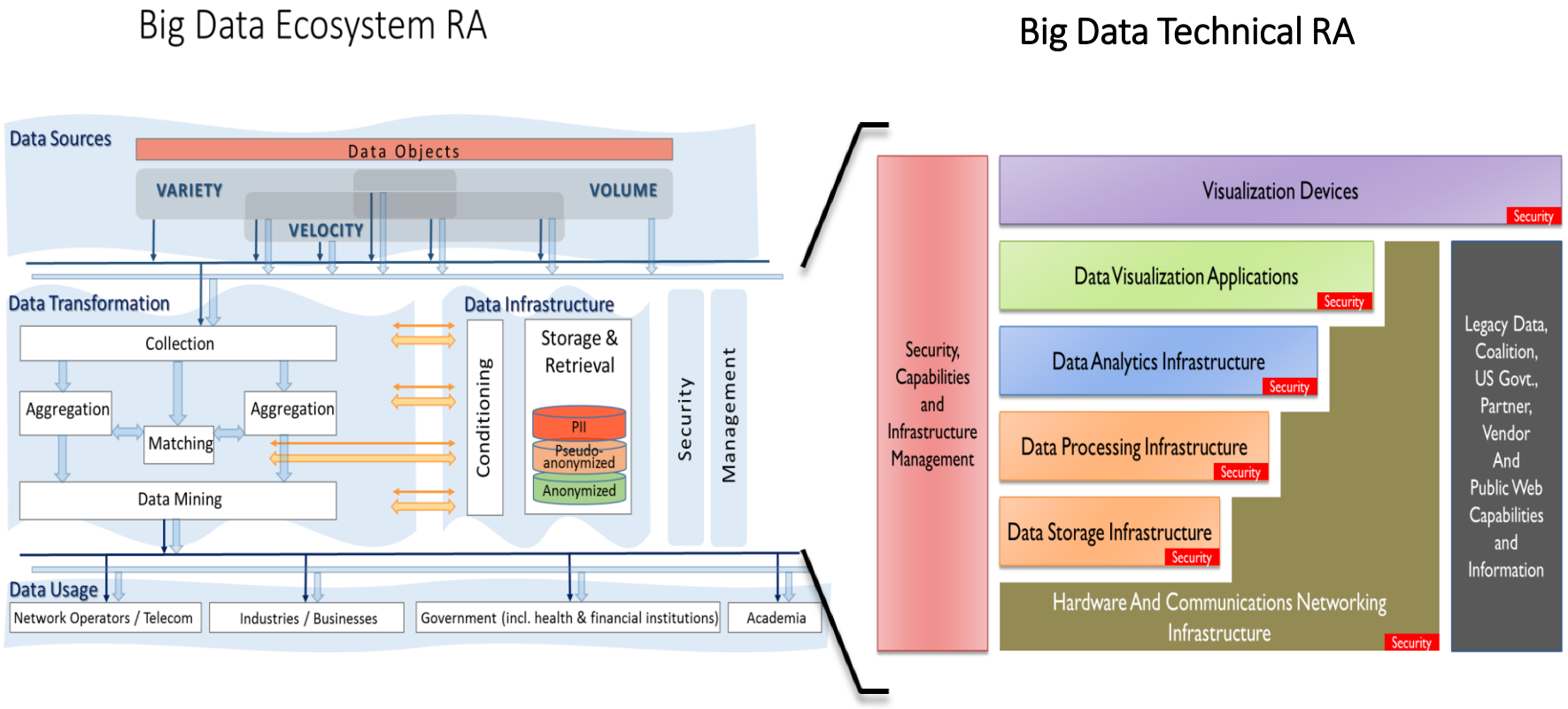
BIG DATA General Reference Architecture:
Deployment Viewpoint

NOTE: Looking For Correct Slides

BIG DATA General Reference Architecture:
Deployment Viewpoint

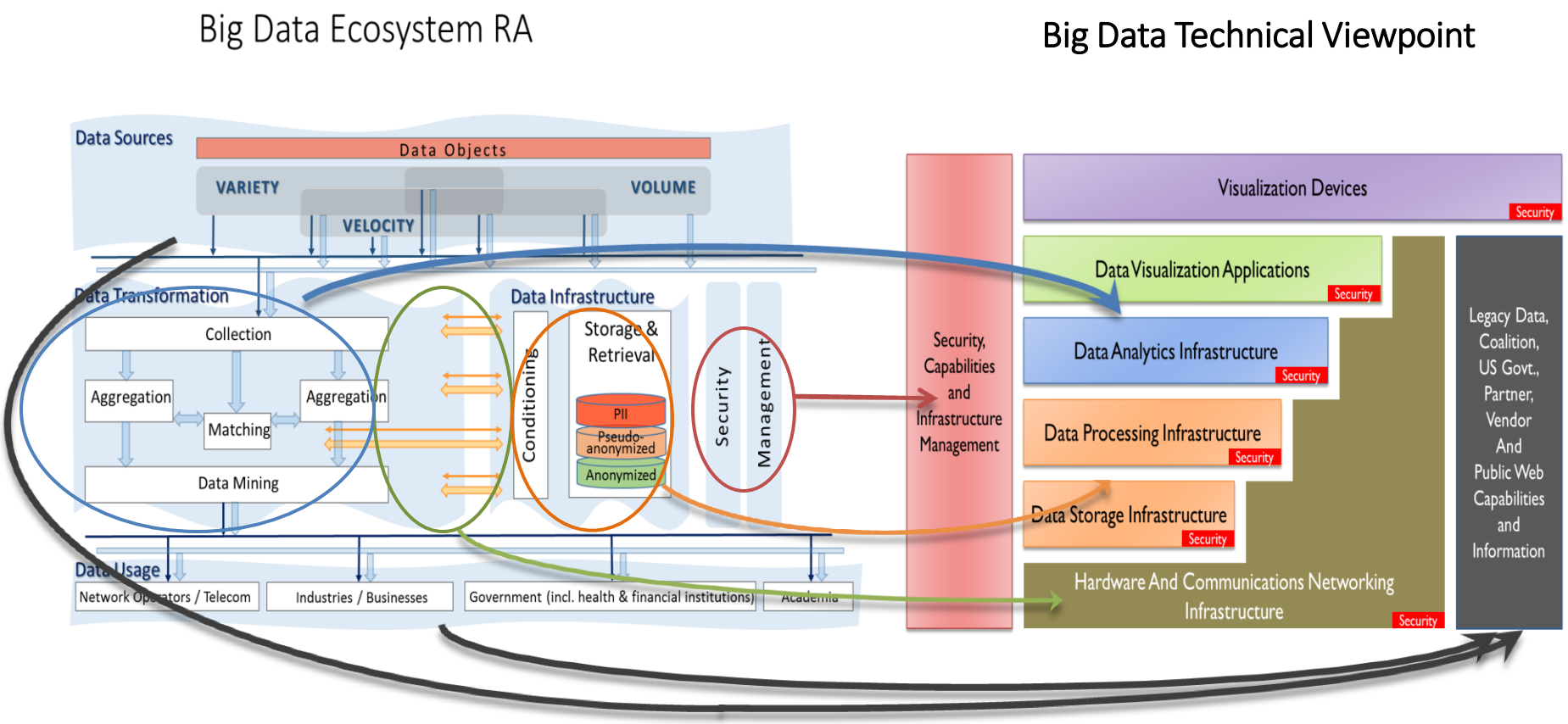
NOTE: Looking For Correct Slides

BIG DATA General Reference Architecture: Eco System RA/Technical RA Capability Alignment



Updated: Added Slide, 07.2013

BIG DATA General Reference Architecture: Eco System RA/Technical RA Capability Mapping



Updated: Added Slide, 07.2013

BIG DATA General Reference Architecture:
Possible Applicable Commercial/Enterprise Functional Standards

- Identity/Security – SAML2, LDAP, PKI, X509, SSL, KMIP
- Authorization – SAML2, VOMS, Shibboleth
- Systems Monitoring – DMTF/CIM, SNMP, ISO X.700-CMIS/CMOT, JMS
- Billing Records -TMF/IPDR
- Cloud Resource Mgt – OGF/OCCI, DMTF/CIMI-OVF, IEEE-P2302(Intercloud RA)
- Grid Resource Mgt – OFG specifications, Globus Speficiations
- Data Management – SNIA/CDMI, OASIS CMIS, OGF specifications
- Storage Management – SNIA/SMIS
- Storage Interface – OSD, SCSI, SATA, SAS, iSCSI, Ficon
- File Sharing – CIFS, NFS, HTTP, WebDAV, SCP, S/FTP
- Service Protocols – OMG CORBA, REST, SOAP, SOA
- Application Configuration Deployments – OASIS TOSCA
- Infrastructure Configuration Deployments – DMTF CIM
- Data Services – OASIS WSDL WSRF, OFG DFDL specifications
- Data Expression – W3C XML, RDF/a, JSON, RSS, Mitre/NIST CEE family
- Document Formats – PDF, HTML, ODF, SMIL, UN/EDIFACT, many others
- Query Languages - SQL, W3C SPARQL, Xquery/Xpath
- Messaging – SNMP, OASIS AMQP, XMPP, ESB
- Service Agreements – OGF GRAAP, WS-Agreement

BIG DATA General Reference Architecture:
Opportunities For New Functional Standards

- What We Know Today, Ten (10) Key Gaps In Standards for BIG DATA Capabilities

1. Information/Data Interoperability Interface Specification (information structure/translation)(increase data utilization)
2. Information Confidence Grading Specification (trust results)
3. RESTful Cloud Object Management Interface Specification (to drive other new interface specifications)
4. Common Catalog Interface Specification – Searchable Capabilities, Services, Applications, Information, Data (profiles)
5. RESTful URI Search/Query Interface (CDR work?) (reduce dev/ops costs, increase deployment options)
6. Data Virtualization Interface Specification (reduce dev/ops costs, increase deployment options)
7. Infrastructure Management Harmonization Interface Spec. (reduce mgt costs, policy based, autonomic data center mgt)
8. Cloud PAAS/SAAS Management Interface Specification (for workload mgt, improved security)
9. Compute/Data Resource Confidentiality/Authorization Interface Specification (system security)
10. Natural Language Query Specification (extend info harvesting to imaging/video, integrated redaction)

BIG DATA :
Reference Architecture Conclusion

- Proposes an Approach for a Technology Agnostic, General Reference Architecture that provides guidance for Delivering BIG DATA Applications Reference Architectures, Implementation Architectures And Capabilities Road Maps
- Identifies "Capabilities to Functional Mapping" and "Functions to Standards" and a Formalized *General Reference Architecture Document* as Areas of Study
- Provides a Set of General Capabilities Supporting Commercial and Program Agnostic, Defense/IC End User Expectations.
- Identifies Potential Functional Standards To Accelerate Development of Commoditized Infrastructure and Operation Optimizations
- Identifies Opportunities for Additional Standards Increasing Deployment Options and Drive Cost Savings