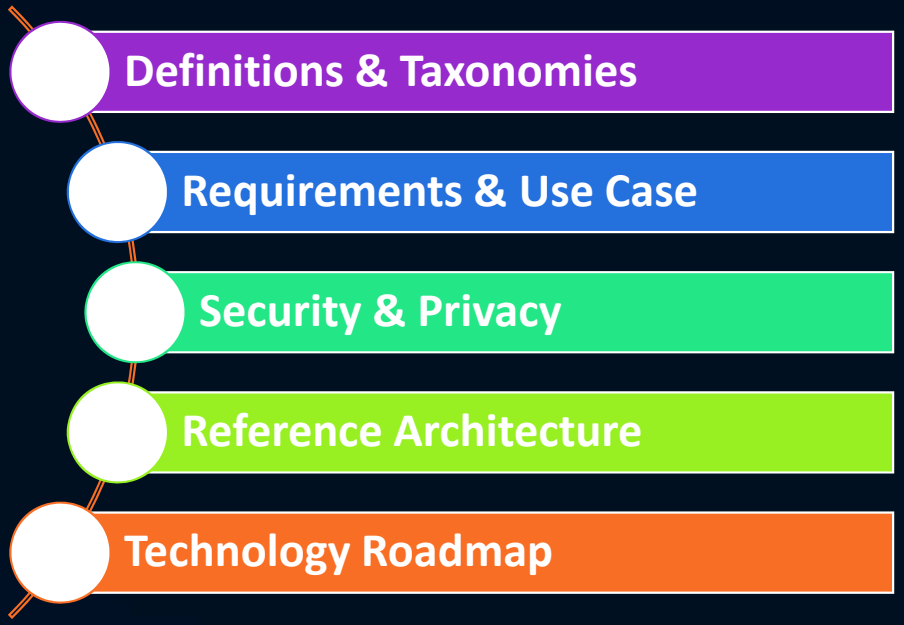
NIST Big Data Public Working Group & Standardization Activities

Wo Chang, NIST, wchang@nist.gov
Robert Marcus, ET-Strategies
Chaitanya Baru, UC San Diego
http://bigdatawg.nist.gov

December 4, 2013

Agenda

- Why Big Data? Why NIST?
- NBD-PWD Charter and Deliverables
- Overall Work Plan
- Subgroup Charter and Deliverables
- Next Steps / Future Activities
- ISO/IEC JTC 1 Big Data Study Group



Why Big Data? Why NIST?

- **Why Big Data?** There is broad agreement among commercial, academic, and government leaders about the remarkable potential of "Big Data" to spark innovation, fuel commerce, and drive progress.

- **Why NIST?** **(a)** Recommendation from January 15 -- 17, 2013 Cloud/Big Data Forum and **(b)** A lack of consensus on some important, fundamental questions is confusing potential users and holding back progress. Questions such as:

  - *What are the attributes that define Big Data solutions?*
  - *How is Big Data different from traditional data environments and related applications?*
  - *What are the essential characteristics of Big Data environments?*
  - *How do these environments integrate with currently deployed architectures?*
  - *What are the central scientific, technological, and standardization challenges needed to accelerate the deployment of robust Big Data solutions?*

Why Big Data? Why NIST?

- **Why Big Data?** There is broad agreement among commercial, academic, and government leaders about the remarkable potential of "Big Data" to spark innovation, fuel commerce, and drive progress.

- **Why NIST?** **(a)** Recommendation from January 15 -- 17, 2013 Cloud/Big Data Forum and **(b)** A lack of consensus on some important, fundamental questions is confusing potential users and holding back progress. Questions such as:

NBD-PWG is being launched to address these questions and is charged to develop consensus definitions, taxonomies, secure reference architecture, and technology roadmap for Big Data that can be embraced by all sectors.

NBD-PWG: Charter and Deliverables

Charter (M0001)

The focus of the (NBD-PWG) is to form a community of interest from industry, academia, and government, with the goal of developing a consensus definitions, taxonomies, secure reference architectures, and technology roadmap. The aim is to create vendor-neutral, technology and infrastructure agnostic deliverables to enable big data stakeholders to pick-and-choose best analytics tools for their processing and visualization requirements on the most suitable computing platforms and clusters while allowing value-added from big data service providers and flow of data between the stakeholders in a cohesive and secure manner.

Deliverables - Working Drafts for:

1. Big Data Definitions
2. Big Data Taxonomies
3. Big Data Requirements & Use Cases
4. Big Data Security & Privacy Requirements
5. Architectures Survey
6. Big Data Reference Architecture
7. Big Data Security & Privacy Architecture
8. Big Data Technology Roadmap

LAUNCH DATE: *June 26, 2013*

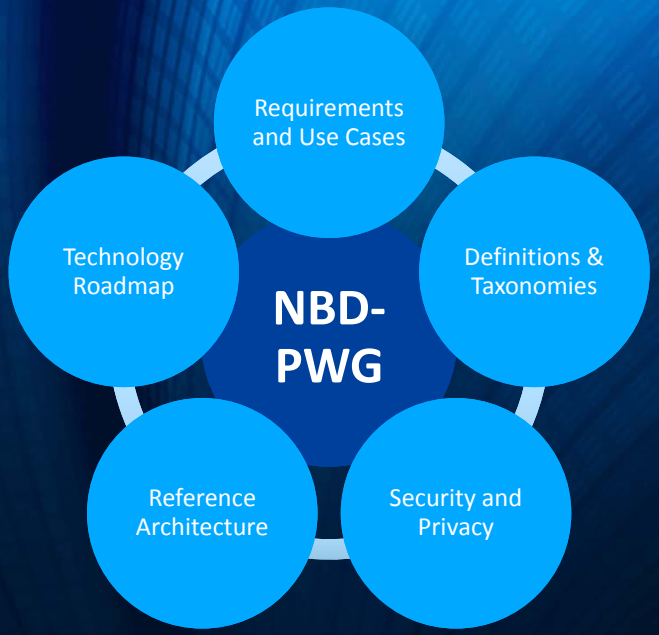
TARGET DATE: *September 27, 2013*

NIST Big Data Public Working Group and Subgroups Work Plan

Week	Def. & Tax.	Requirements	Sec. & Privacy	Ref. Arch	Tech. Roadmap
June 26	NBD-PWG (13:00PM – 15:00PM) Kick-off Meeting				
July 3	NBD-PWG (13:00PM – 15:00PM) Establish Subgroups with Co-Chairs, Subgroups Charter, Overall OWG direction				
July 8 - 12	Mondays 10:00AM – 12:00PM	Tuesdays 10:00AM – 12:00PM	Wednesdays 10:00AM – 12:00PM	Thursdays 10:00AM – 12:00PM	Fridays 10:00AM – 12:00PM
July 15 – 19	Definitions & Characteristics	Collect general use cases, identify requirements	Collect security and privacy use cases,	Analyze use cases from Reqs. & Sec. subgroups	Vision Characteristics & Def.
July 22 – 26	Tax.: Roles, activities, components & subcomp.	Categorize reqs., Identify missing reqs.	Identify requirements	Create conceptual model, identify actors,	Taxonomies Roles & Activities
July 24	NBD-WG (13:00PM – 15:00PM) **Subgroups report: Sharing and brainstorming results**				
July 29 – Aug. 2				Identify usage scenarios, iden. Implement. Scenarios	Use cases & scenarios Ref. Architecture
Aug 5 – 9				Create ref. architecture	Standards & Activities Gap Analysis
Aug 12 – 16					Standardization Priorities ??? Strategy of Adoption
Aug 19 – 23					Strategy of Implement. Resourcing
Aug. 21	NBD-WG (13:00PM – 15:00PM) **Subgroups report: Present and Discuss Working Draft Outline**				
Aug. 26 - 30					Recommendations
Sept. 2 – 6					
Sept. 4	NBD-WG (13:00PM – 15:00PM) **Subgroups report: Present and Discuss Rough Draft**				
Sep 9 – 13					
Sep 16 - 20					
Sep 23 – 27					
Sep 25	NBD-WG (13:00PM – 15:00PM) **Subgroups report: Present and Discuss Final Draft**				
Sep 30	Big Data Workshop, NIST **- Deliverables Presentation & Discussion** **- Breakout Sessions by Subgroups** - Announcement for Next Steps				

SUBGROUPS

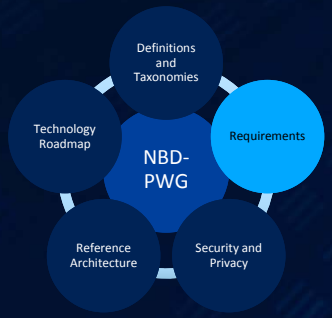
AND THEIR SCOPES AND DELIVERABLES



SUBGROUP

Requirements & Use Cases

Geoffrey Fox, U. Indiana
Joe Paiva, VA
Tsegereda Beyene, Cisco



Scope (M0020)

The focus is to form a community of interest from industry, academia, and government, with the goal of developing a consensus list of Big Data requirements across all stakeholders. This includes gathering and understanding various use cases from diversified application domains.

Tasks

- Gather input from all stakeholders regarding Big Data requirements.
- Analyze/prioritize a list of challenging general requirements that may delay or prevent adoption of Big Data deployment
- Develop a comprehensive list of Big Data requirements

Requirements and Use Case Subgroup

Key documents:

M0105 – Use Cases

M0125 – Requirements

Mo152 – Working Draft

Use Case Template:

1. Goals, Description
2. Data Characteristics, Data Types
3. Data Analytics
4. Current Solutions
5. Security & Privacy
6. Lifecycle Management & Data Quality
7. System Management & Other Issues

NBD (NIST Big Data) Requirements WG Use Case Template Aug 11 2013

Use Case Title		
Vertical (area)		
Author/Company/Email		
Actors/Stakeholders and their roles and responsibilities		
Goals		
Use Case Description		
Current Solutions	Compute(System)	
	Storage	
	Networking	
	Software	
Big Data Characteristics	Data Source (distributed/centralized)	
	Volume (size)	
	Velocity (e.g. real time)	
	Variety (multiple datasets, mashup)	
	Variability (rate of change)	
Big Data Science (collection, curation, analysis, action)	Veracity (Robustness Issues, semantics)	
	Visualization	
	Data Quality (syntax)	
	Data Types	
	Data Analytics	
Big Data Specific Challenges (Gaps)		
Big Data Specific Challenges in Mobility		
Security & Privacy Requirements		
Highlight issues for generalizing this use case (e.g. for ref. architecture)		
More Information (URLs)		
Note: <additional comments>		

Note: No proprietary or confidential information should be included

Requirements and Use Case Subgroup

51 Use Cases Received (http://bigdatawg.nist.gov/usecases.php)

1. **Government Operations (4):** National Archives & Records Administration, Census Bureau
2. **Commercial (8):** Finance in Cloud, Cloud Backup, Mendeley (Citations), Netflix, Web Search, Digital Materials, Cargo shipping (e.g. UPS)
3. **Defense (3):** Sensors, Image Surveillance, Situation Assessment
4. **Healthcare & Life Sciences (10):** Medical Records, Graph & Probabilistic Analysis, Pathology, Bio-imaging, Genomics, Epidemiology, People Activity Models, Biodiversity
5. **Deep Learning & Social Media (6):** Driving Car, Geolocate Images, Twitter, Crowd Sourcing, Network Science, NIST Benchmark Datasets
6. **The Ecosystem for Research (4):** Metadata, Collaboration, Language Translation, Light Source Experiments
7. **Astronomy & Physics (5):** Sky Surveys, Large Hadron Collider at CERN, Belle Accelerator II (Japan)
8. **Earth, Environmental & Polar Science (10):** Ice Sheet Scattering, Earthquake, Ocean, Earth Radar Mapping, Climate Simulation, Atmospheric Turbulence, Subsurface Biogeochemistry, AmeriFlux &FLUXNET gas sensors
9. **Energy (10):** Smart Grid

Requirements and Use Case Subgroup

Step 1 **Extract requirements and map to reference architecture based on application characteristics:**

a. **Data sources** (data size, file formats, rate of growth, at rest or in motion, etc.)

b. **Data lifecycle management** (curation, conversion, quality check, pre-analytic processing, etc.)

c. **Data transformation** (data fusion/mashup, analytics)

d. **Capability infrastructure** (software tools, platform tools, hardware resources like storage and networking)

e.

f.

g.

437 requirements were extracted from 51 Use Cases
35 aggregated general requirements divided into 7 categories

Step 2 **Aggregate all specific requirements into high-level generalized requirements which are vendor-neutral and technology agnostic**

Requirements and Use Case Subgroup

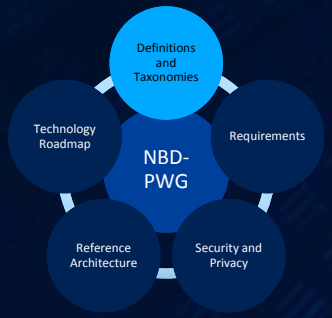
Part of Property Summary Table

No.	Use Case	Volume	Velocity	Variety	Software	Analytics
24	M0173 Social Contagion Modeling for Planning	10s of TB per year	During social unrest events, human interactions and mobility leads to rapid changes in data; e.g., who follows whom in Twitter.	Data fusion a big issue. How to combine data from different sources and how to deal with missing or incomplete data?	Specialized simulators, open source software, and proprietary modeling environments. Databases.	Models of behavior of humans and hard infrastructures, and their interactions. Visualization of results
25	M0141 Biodiversity and LifeWatch	N/A	Real time processing and analysis in case of the natural or industrial disaster	Rich variety and number of involved databases and observation data	RDMS	Requires advanced and rich visualization
26	M0136 Large-scale Deep Learning	Current datasets typically 1 to 10 TB. Training a self-driving car could take 100 million images.	Much faster than real-time processing is required. For autonomous driving need to process 1000's high-resolution (6 megapixels or more) images per second.	Neural Net very heterogeneous as it learns many different features	In-house GPU kernels and MPI-based communication developed by Stanford. C++/Python source.	Small degree of batch statistical pre-processing; all other data analysis is performed by the learning algorithm itself.
27	M0171 Organizing large-scale image collections	500+ billion photos on Facebook, 5+ billion photos on Flickr.	over 500M images uploaded to Facebook each day	Images and metadata including EXIF tags (focal distance, camera type, etc),	Hadoop Map-reduce, simple hand-written multithreaded tools (ssh and sockets for communication)	Robust non-linear least squares optimization problem. Support Vector Machine
28	M0160 Truthy	30TB/year compressed data	Near real-time data storage, querying & analysis	Schema provided by social media data source. Currently using Twitter only. We plan to expand	Hadoop IndexedHBase & HDFS. Hadoop, Hive, Redis for data management. Python:	Anomaly detection, stream clustering, signal classification and online-learning; Information diffusion,

SUBGROUP

Definitions & Taxonomies

Nancy Grady, SAIC
Natasha Balac, SDSC
Eugene Luster, R2AD



Scope (M0018)

The focus is to gain a better understanding of the principles of Big Data. It is important to develop a consensus-based common language and vocabulary terms used in Big Data across stakeholders from industry, academia, and government. In addition, it is also critical to identify essential actors with roles and responsibility, and subdivide them into components and sub-components on how they interact/ relate with each other according to their similarities and differences.

Tasks

- For Definitions: Compile terms used from all stakeholders regarding the meaning of Big Data from various standard bodies, domain applications, and diversified operational environments.
- For Taxonomies: Identify key actors with their roles and responsibilities from all stakeholders, categorize them into components and subcomponents based on their similarities and differences
- Develop Big Data Definitions and taxonomies documents

Definitions and Taxonomies Subgroup (M0024, M0142)

Key documents:

M0024 – Ongoing Discussion

M0142 – Working Draft

Big Data Definitions, v1 (Developed from Jan. 15-17, 2013 NIST Cloud/Big Data Workshop)

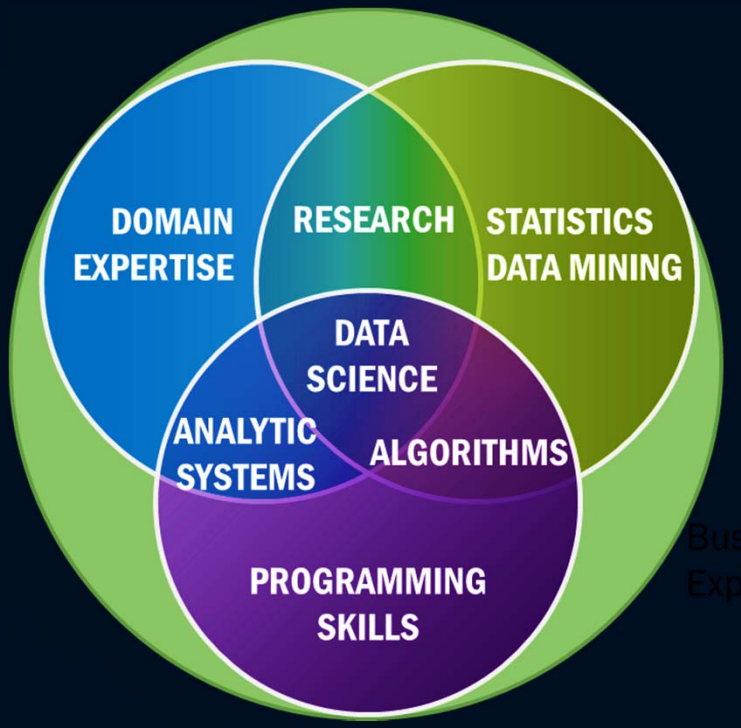
Big Data refers to digital data volume, velocity and/or variety that:

- **Enable** novel approaches to frontier questions previously inaccessible or impractical using current or conventional methods; and/or
- **Exceed** the storage capacity or analysis capability of current or conventional methods and systems; and
- **Differentiates** by storing and analyzing population data and not sample sizes.

Definitions and Taxonomies Subgroup (M0024, M0142)

- **Data Science** is the extraction of actionable knowledge directly from data through a process of discovery, hypothesis, and analytical hypothesis analysis.

- **Data Scientists** is a practitioner who has sufficient knowledge of the overlapping regimes of expertise in business needs, domain knowledge, analytical skills and programming expertise to manage the end-to-end scientific method process through each stage in the Big Data lifecycle (through action) to deliver value.



Definitions and Taxonomies Subgroup (M0024, M0142)

Big Data Taxonomies (M0202)

Actors

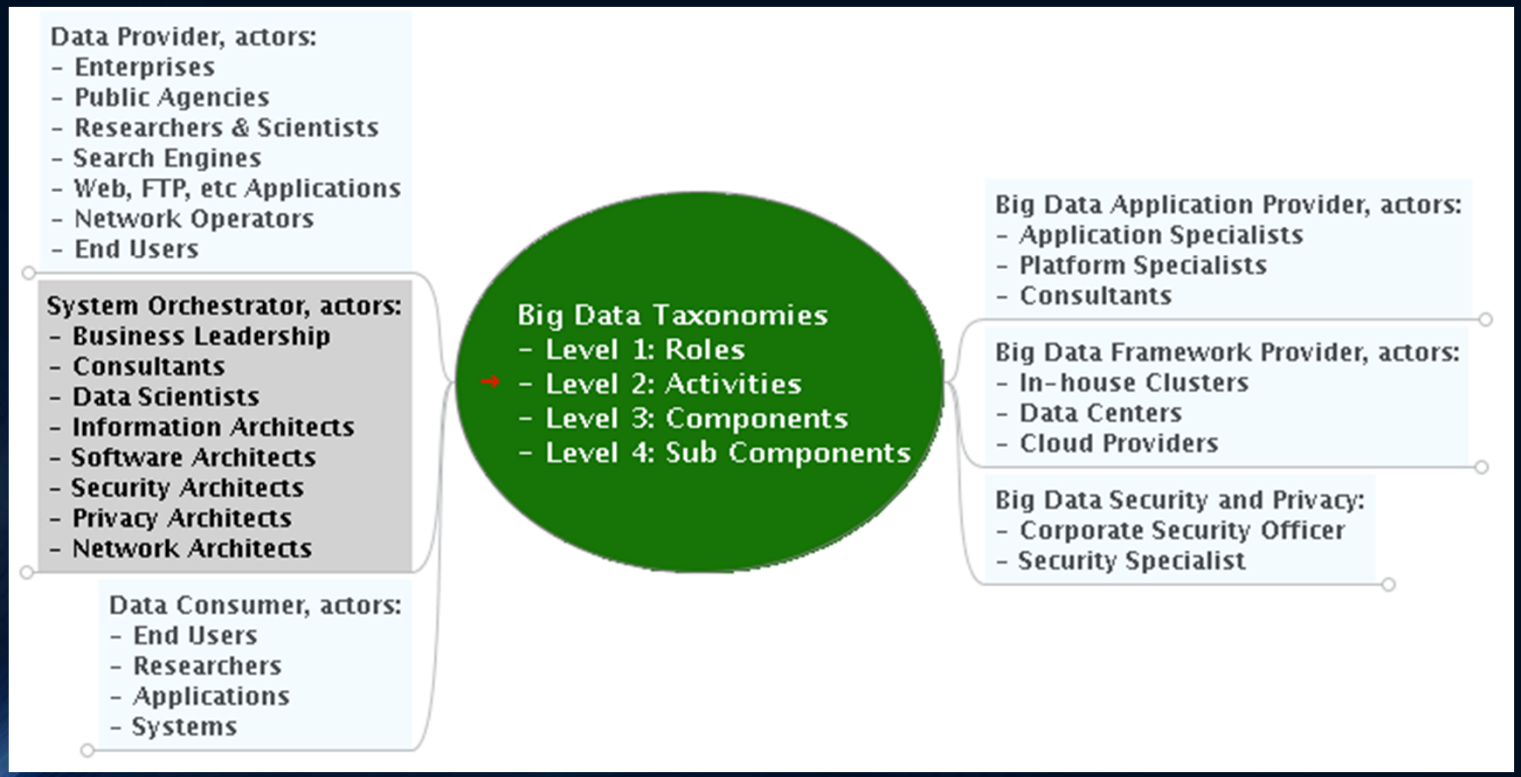
1. Sensors
2. Applications
3. Software agents
4. Individuals
5. Organizations
6. Hardware resources
7. Service abstractions

System Roles

1. **Data Provider**
   > Makes available data internal and/or external to the system
2. **Data Consumer**
   > Uses the output of the system
3. **System Orchestrator**
   > Governance, requirements, monitoring
4. **Big Data Application Provider**
   > Instantiates application
5. **Big Data Framework Provider**
   > Provides resources

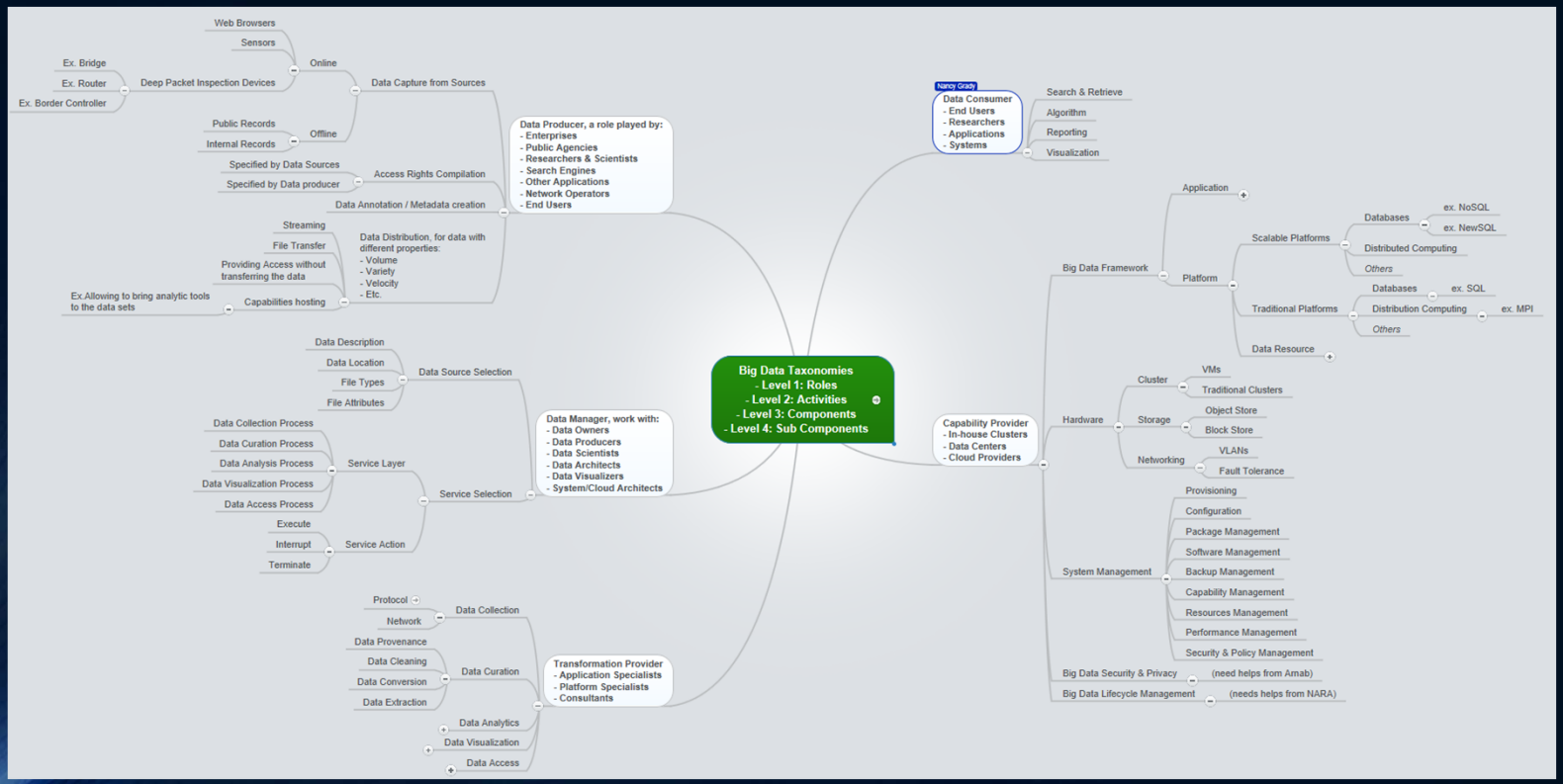
Definitions and Taxonomies Subgroup (M0024, M0142)

Big Data Taxonomies (M0202)



Definitions and Taxonomies Subgroup (M0024, M0142)

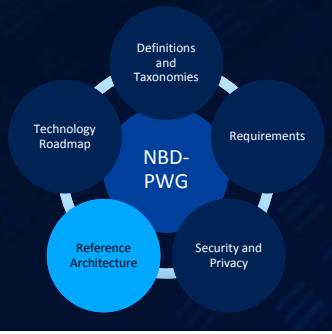
Big Data Taxonomies (M0202)



SUBGROUP

Reference Architecture

Orit Levin, Microsoft
James Ketner, AT&T
Don Krapohl,
Augmented Intelligence



Scope (M0021)

The focus is to form a community of interest from industry, academia, and government, with the goal of developing a consensus-based approach to orchestrate vendor-neutral, technology and infrastructure agnostic for analytics tools and computing environments. The goal is to enable Big Data stakeholders to pick-and-choose technology-agnostic analytics tools for processing and visualization in any computing platform and cluster while allowing value-added from Big Data service providers and the flow of the data between the stakeholders in a cohesive and secure manner.

Tasks

- Gather and study available Big Data architectures representing various stakeholders, different data types,' use cases, and document the architectures using the Big Data taxonomies model based upon the identified actors with their roles and responsibilities.
- Ensure that the developed Big Data reference architecture and the Security and Privacy Reference Architecture correspond and complement each other.

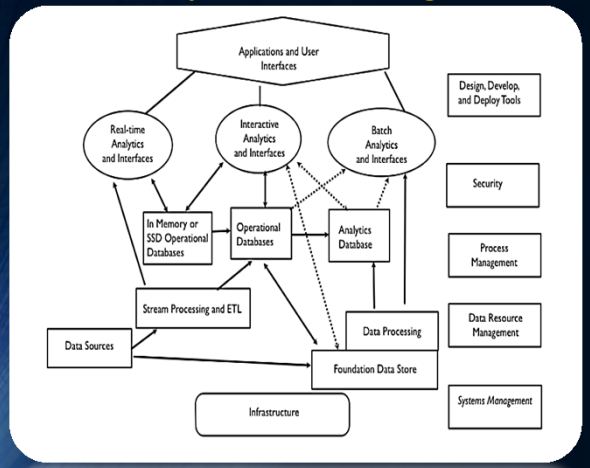
Reference Architecture Subgroup

Key documents:

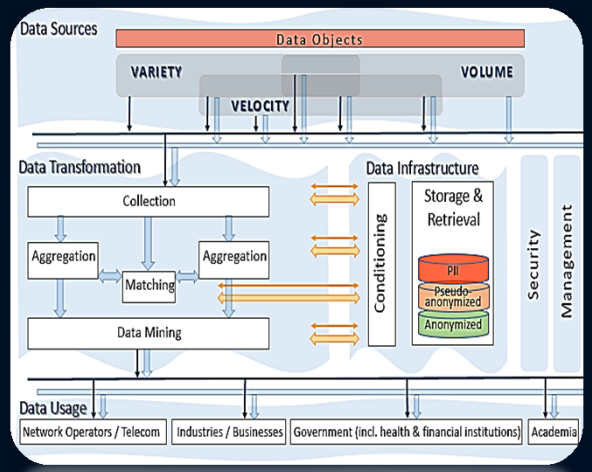
M0151 – White Paper

M0123 – Working Draft

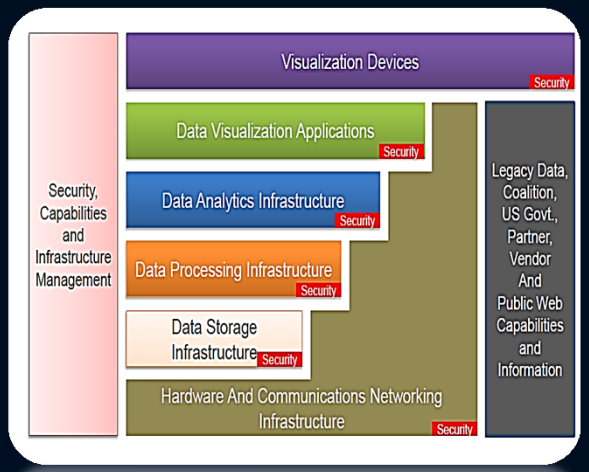
M0039 | Data Processing Flow



M0017 | Data Transformation Flow

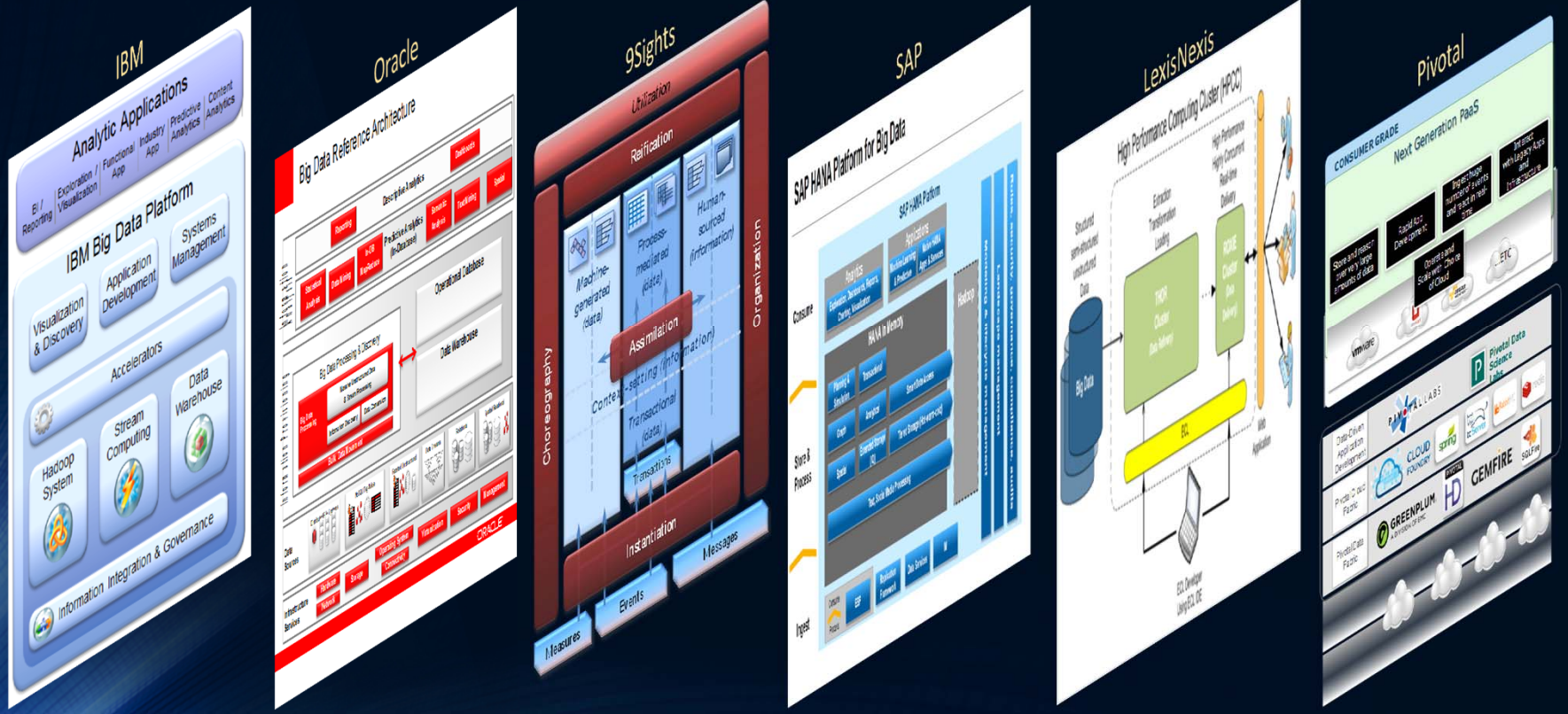


M0047 | IT Stack



Reference Architecture Subgroup

Vendors Big Data Architectures



Reference Architecture Subgroup

What the Baseline Big Data RA

IS

- ✓ A superset of the "traditional data" system
- ✓ A representation of a vendor-neutral & 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

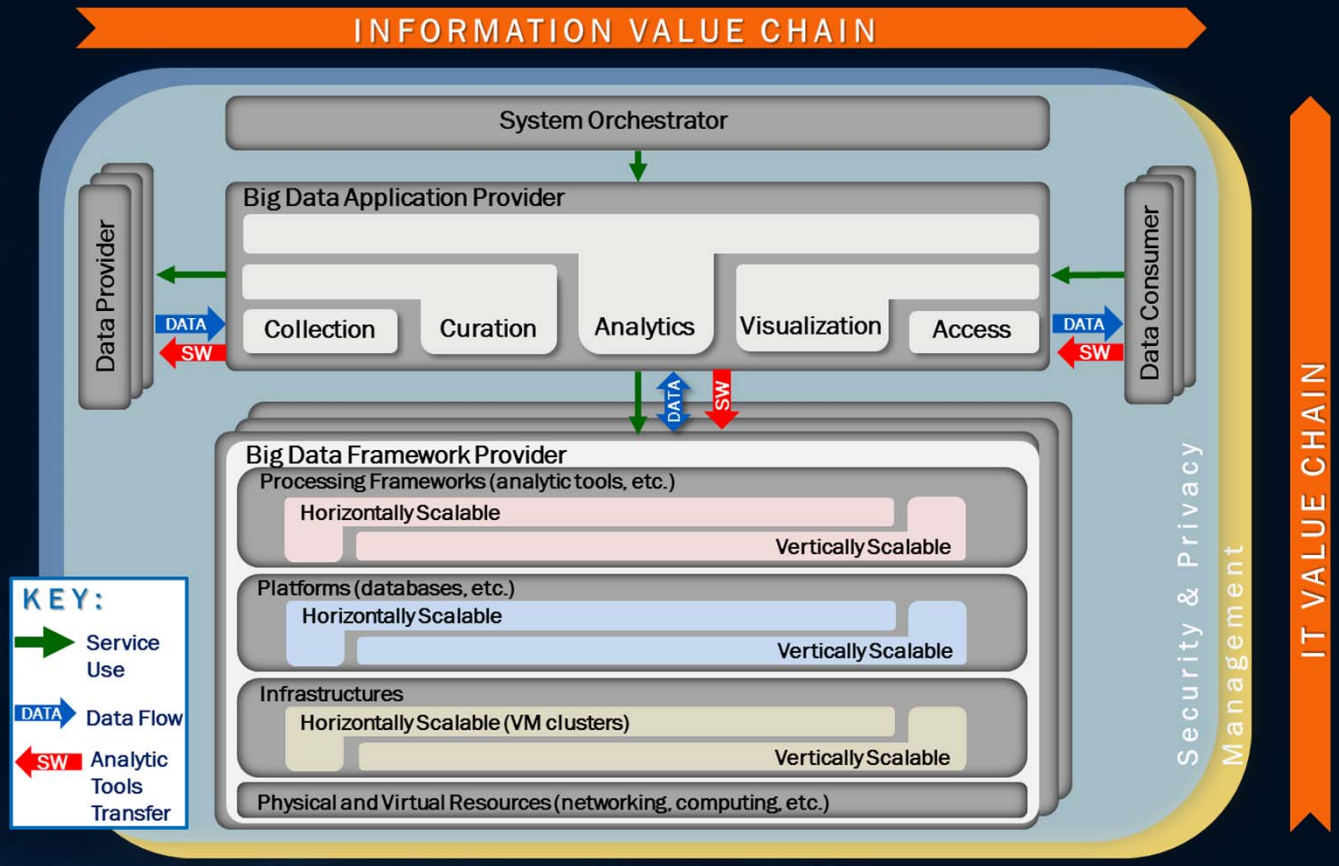
IS NOT

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

All of the above will be developed in the next stage in the context of specific use cases.

Reference Architecture Subgroup

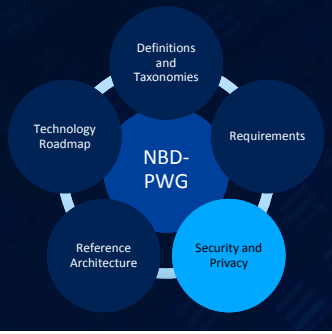
RA Diagram



SUBGROUP

Security & Privacy

Arnab Roy, CSA/Fujitsu Nancy Landreville, U. MD Akhil Manchanda, GE



Scope (M0019)

The focus is to form a community of interest from industry, academia, and government, with the goal of developing a consensus secure reference architecture to handle security and privacy issues across all stakeholders. This includes gaining an understanding of what standards are available or under development, as well as identifies which key organizations are working on these standards.

Tasks

- Gather input from all stakeholders regarding security and privacy concerns in Big Data processing, storage, and services.
- Analyze/prioritize a list of challenging security and privacy requirements that may delay or prevent adoption of Big Data deployment
- Develop a Security and Privacy Reference Architecture that supplements the general Big Data Reference Architecture

Security and Privacy Subgroup

Key documents:

Google Doc – Ongoing Discussion

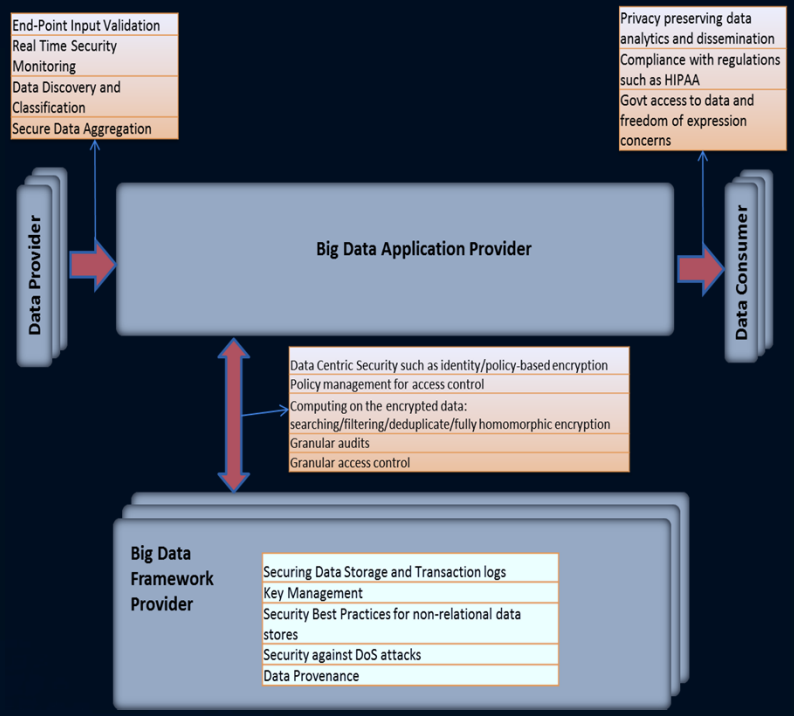
M0110 – Requirements Working Draft

M0xxx – Architecture & Taxonomies

Requirements Scope
- Infrastructure Security
- Data Privacy
- Data Management
- Integrity & Reactive Security

Requirements – Use Cases Studied
- Retail (consumer)
- Healthcare
- Media
- Government
- Marketing

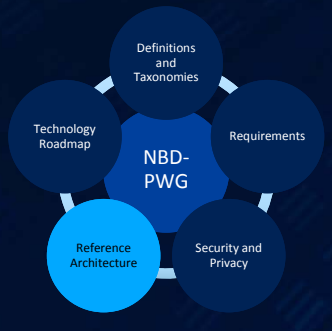
Architecture & Taxonomies
- Privacy
- Provenance
- System Health



SUBGROUP

Technology Roadmap

Carl Buffington, USDA/Vistronix
Dan McClary, Oracle
David Boyd, Data Tactic



Scope (M0022)

The focus is to form a community of interest from industry, academia, and government, with the goal of developing a consensus vision with recommendations on how Big Data should move forward by performing a good gap analysis through the materials gathered from all other NBD subgroups. This includes setting standardization and adoption priorities through an understanding of what standards are available or under development as part of the recommendations.

Tasks

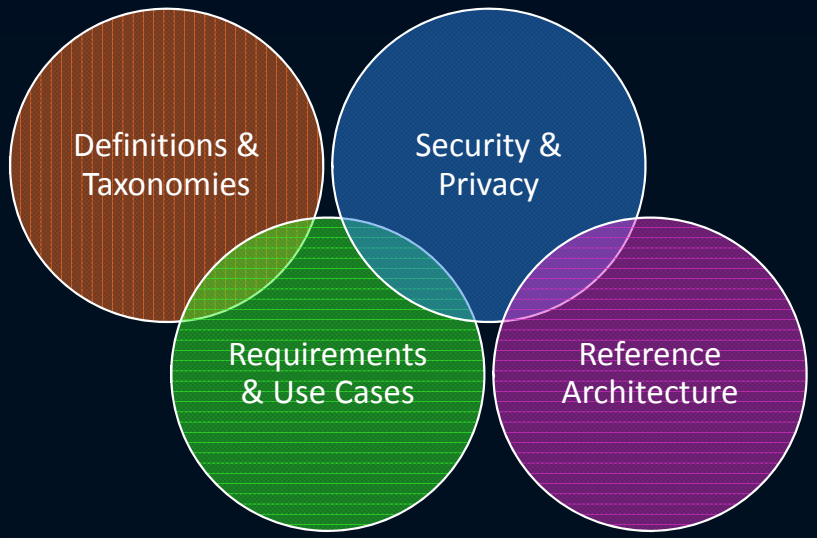
- Gather input from NBD subgroups and study the taxonomies for the actors' roles and responsibility, use cases and requirements, and secure reference architecture.
- Gain understanding of what standards are available or under development for Big Data
- Perform a thorough gap analysis and document the findings
- Identify what possible barriers may delay or prevent adoption of Big Data
- Document vision and recommendations

Technology Roadmap Subgroup

Key document:

M0087 – Working Draft

- Inputs from other subgroups
- Potential Standards Group with Big Data-related activities (M0035)
- Capabilities & Technology Readiness
- Big Data Decision Framework
- Big Data Mapping & Gap Analysis
- Big Data Strategies



Subgroups Working Draft Outline

Contact: bigdatainfo@nist.gov

Website: http://bigdatawg.nist.gov

Join NBD-PWG: http://bigdatawg.nist.gov/newuser.php

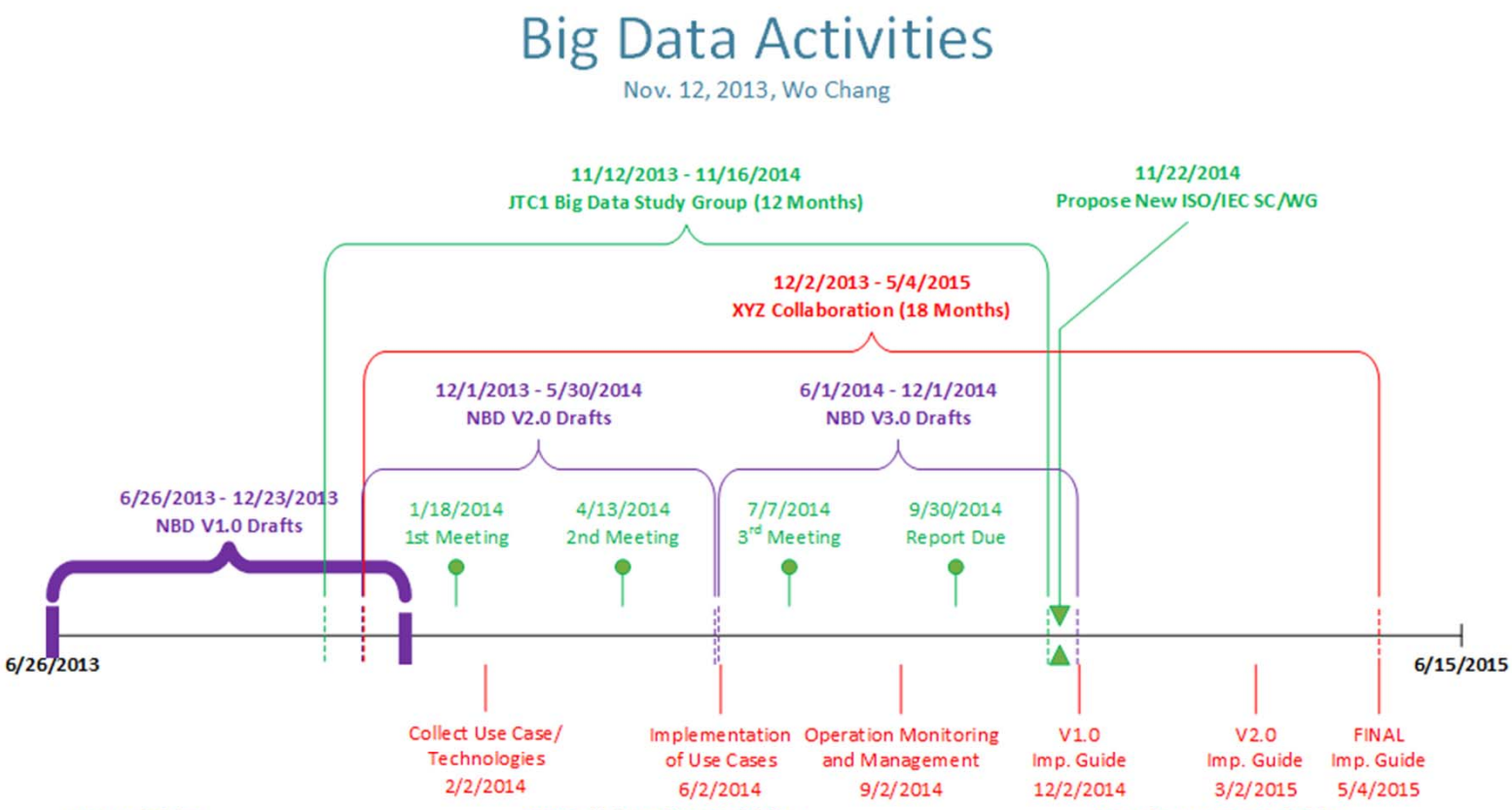
Documents: http://bigdatawg.nist.gov/show_InputDoc.php

Working Drafts

- Big Data Definitions & Taxonomies (M0142)
- Big Data Requirements (M0245)
- Big Data Security & Privacy Requirements (M0110)
- Big Data Architectures White Paper Survey (M0151)
- Big Data Reference Architectures (M0226)
- Big Data Security & Privacy Reference Architecture (M0110)
- Big Data Technology Roadmap (M0087)

NIST Big Data Workshop Slides: http://bigdatawg.nist.gov/workshop.php

Next Steps & Future Activities



NBD Activities
- V1.0, Reference Architecture
- V2.0, Architecture Interfaces
- V3.0, Analytics Tool & Application

Deliverables
- NBD V1.0, Dec. 2013
- NBD V2.0, May, 2014
- NBD V3.0, Dec., 2014

XYZ Collaboration Activities
- Collect unique use cases
- Explore related technologies
- Testbed implementation
- Explore monitoring/management tools

Deliverables
- Best Practice Imp. Guide V1.0, Dec. 2014
- Best Practice Imp. Guide V2.0, Mar. 2015
- Best Practice Imp. Guide Final, May, 2015

JTC1 BigData SG Activities
- Big Data Architecture
- Big Data Security & Privacy
- Big Data Analytics
- Big Data Management
- Big Data Applications & Tools

Deliverables
- US meeting, NIST SP, Feb., 2014
- Europe meeting, NIST SP, May, 2014
- Asia meeting, NIST SP, Aug, 2014
- Big Data SG Report, Sep. 30, 2014

ISO/IEC JTC 1 Big Data Study Group

Convener: Wo Chang, NIST

Terms and References:

1. Survey the existing ICT landscape for key technologies and relevant standards/models/studies/use cases and scenarios for Big Data from JTC 1, ISO, IEC and other standards setting organizations
2. Identify key terms and definitions commonly used in the area of Big Data
3. Assess the current status of Big Data standardization market requirements, identify standards gaps, and propose standardization priorities to serve as a basis for future JTC 1 work
4. Provide a report with recommendations and other potential deliverables to the 2014 JTC1 Plenary

Focus Areas:

1. Big Data Architecture / Infrastructure
2. Big Data Security & Privacy
3. Big Data Analytics
4. Big Data Applications & Tools
5. Big Data Management

Meetings (3 face-t0-face + teleconferences):

Format – 4 days (2 days workshop + 2 day meetings)

January – U.S. ✶ April – Europe ✶ July – Asia

Additional Notes:

1. Workshop papers will go into NIST Special Publication
2. High-quality papers may go into ACM Conference Proceedings (in process)
3. Report due in September, 2014

Questions?

PLEASE CONTACT: WCHANG@NIST.GOV