Presentation Overview

- Volume Presentation Outline
- **Volume 1, Definitions (Nancy Grady, SAIC)**
- Volume 2, BD Taxonomies (Nancy Grady, SAIC)
- Volume 3, Use Cases and General Requirements (Geoffrey Fox, Indiana University)
- Volume 6, Reference Architecture (David Boyd, InCadence Corp.)
- Volume 4, Security and Privacy (Arnab Roy, Fujitsu; Mark Underwood, AVP, Strategic Initiatives, Controls and Countermeasures)
- Volume 8, Reference Architecture Interface (Gregor von Laszewski, Indiana University)
- Reference Architecture Software Implementation Environment and Demonstration (Gregor von Laszewski, Indiana University)
- Volume 7, Standards Roadmap (Russell Reinsch, Center for Government Interoperability)
- Volume 9, Adoption and Modernization (Russell Reinsch, Center for Government Interoperability)
NBDIF Volume Overview

Vol. 1 BD Definitions
Defines common language

Vol. 2 BD Taxonomies
Hierarchy of NBDRA components

Vol. 3 Use Cases & Vol. 5 Arch Survey
Info gathered; requirements extracted

Vol. 4 S&P
Interwoven topics of S&P examined

Vol. 6 NBDRA
Developed NBDRA

Vol. 7 Standards Roadmap
Examine standards wrt NBDRA

Vol. 8 NBDRA Interfaces
Implementation of NBDRA

Vol. 9 Adoption & Modernization
Volume Presentation Outline

- For each volume
  - Scope of the volume
  - Brief recap of version 1
  - Highlights of version 2 accomplishments
  - Summary of version 2 areas needing contributions
  - Topics that could be considered for version 3
Volume 1, Definitions

Document Scope

• Define terminology used in community
• Define terminology used in the other volumes of the NBDIF
• Definition of Big Data, Data Science, and related terms
• Narrative description to add conceptual framework around Big Data terminology
• Provides vocabulary to clarify discussions surrounding Big Data
• Audience anyone who is:
  – New to Big Data to understand concepts
  – Want to be compliant with a common vocabulary
  – Need to evaluate vendor concepts
Volume 1, Definitions

Version 1 Overview

• Big Data and Data Science Definitions
  – **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.
  – **Data science** is the extraction of useful knowledge directly from data through a process of discovery, or of hypothesis formulation and hypothesis testing.
  – Comparison to range of Big Data definitions that have been published
Big Data Features – clarify what is in scope
  - Data types and metadata (not new)
  - Data records (Non-Relational Models *not NoSQL*)
  - Datasets
    - Distributed storage
    - Distributed computing
    - Resource Negotiation
    - Datasets in Motion (streaming data)
    - Data Science Lifecycle Model
  - Big Data Analytics (looking at V’s)
Volume 1, Definitions

Version 1 Overview (cont)

• **Areas introduced but not covered**
  – Big Data Metrics
  – Big Data Security and Privacy
  – Data Governance
  – Big Data Engineering Patterns
Volume 1, Definitions

Version 2 Accomplishments

• Big Data
  – Volume, Velocity, Variety, Variability

• Expanded discussion of Big Data Engineering Frameworks
  – Horizontal infrastructure scaling
  – Scalable logical data storage
  – Relationship to other technological innovations
    • HPC, Cloud, IoT, Cyber-Physical Systems, Blockchain

• Reorganized the analysis of big data – i.e. Data Science
  – Veracity, Validity, Visualization, Value
  – Metadata, Data Types, Complexity, Latency
  – But not pre-existing cleanliness, completeness, etc
Volume 1, Definitions

Version 2 Accomplishments

• Expand Big Data Science novelty
  – Machine learning
  – Emergent Behavior
  – Data Scientists
  – Benchmarks

• Big Data security and privacy – still summary of Vol 4

• Management groundwork discussion and definitions
  – Orchestration
  – Governance
  – Data Ownership
  – Societal Implications
Volume 1, Definitions

Version 2 Opportunities for Contribution

• **Concurrency** definition and discussion (Section 3.1)
• Enhanced discussion of **HPC** (S3.3.1), **Cloud** (S3.3.2), **IoT** (S3.3.3), **CPS** (S3.3.4), **Blockchain** (S3.3.5)
• **Latency**: describe and relate to Big Data (S4.2.9)
• **Emergent Behavior**: description and relation to Big Data (S4.4)
• **Data cleansing**: describe and relate to Big Data (S4.3.1)
• **Machine learning**: describe and relate to Big Data (S4.3.3)
• **Big Data Management** (S6.0): discuss wrt Big Data and orchestration (S6.1), data governance (S6.2), and data ownership (S6.3)
• Pointers to external materials not covered here in detail
• References to parallel works by others
Possible Version 3 Topics

- Categorization of Relational/NoSQL/NewSQL/etc attributes
  - To assist in implementation comparisons
- Metrics guidance
- Discussion of Visualization
  - Exploratory, Evaluative, Explanatory
  - Augmented Reality and Virtual Reality
- Expansion of Machine Learning/Deep Learning/Artificial Intelligence
- Algorithms and Analytics Frameworks
- Dedicated Languages ???
- Emerging topics - ???
Volume 1, Definitions

Breakout Plan

- Review Version 2 slide of remaining items
  - Do any need so much work they should be deferred to version 3
- What have we missed
- What is not needed or is poorly expressed
- Review of Version 3 slide
  - Anything that should be put in version 2
  - Anything missing