NIST Big Data Public Working Group

Breakout Overview Report

NIST Campus
Gaithersburg, Maryland
June 2, 2017

Breakout 1 – Nancy Grady

Volume 1, Definitions

- Additional Clarification
  - Distinguish fusion from alignment from integrated (variety)
  - Need for retention/deletion policies (governance)
  - Add data sharing discussion (governance)
  - Progression statistics -> data mining -> data science
  - Horizontal vs vertical data scientist (or not a data scientist)
    - Make data scientist an umbrella term?
  - Survey of data scientist definitions
  - Provenance and data citation
- Suggestions
  - Data Acquisition instead of Data Collection

Breakout 1 – Nancy Grady

Volume 1, Definitions (cont.)

- Additions
  - Linked Data and RDF
  - Open Data
  - Expand Machine Learning/Deep Learning (NLP, image, etc)
  - Open Science
  - Data Models (mapping real world)
- Related Topics wrt Big Data (current are mostly infrastructure)
  - Multimedia
  - Streaming data analytics
- Context suggestions - history from digitization
- Pointer – paper “A formal definition of Big Data based on its essential features”

Breakout 1 – Nancy Grady

Volume 2, Big Data Taxonomies

- Expand
  - Metadata at each data level
- Missing
  - Discussion of other Big Data Taxonomies
- Pointers
  - W3C PROV
- Consider
  - Taxonomies vs models (such as layer models)

Breakout 2 – Geoffrey Fox

Volume 3, Use Cases and Requirements

- Additions
  - More use cases including those in the commercial space
  - Will supply latest Version 2 template in next week. Ask us!!
- The use cases can help to understand Issues
  - such as Logically Centralized vs. Explicitly Decentralized data
    - Commercial cloud centralized data
    - Centralized data is powerful but it may not work for some applications due to distributed ownership (e.g., scientific)
  - Pleasingly parallel use cases shown in use cases; implies system architecture
  - Organizational, political challenges with Big Science Big Data adoption
- Simulations as part of the RA scope?
  - Use of simulations to model job platform/profiles, may also help in the scale up of SnP solutions

Breakout 2 – Mark Underwood, Arnab Roy

Volume 4, Security and Privacy

- Possible additional topics
  - Issues with identity management scalability
  - Device registration processes and standardization
  - Issues surrounding system decommissioning were suggested (volume covers life cycle but not much about decommissioning)
  - Service layer added to NBDRA?
  - Key rotation in the discussion of key management
- Consider
  - Health care data provenance
  - Review FIPS 140-2
- No volunteers to draft text from breakout
- Key issues of model based engineering and domain models was neither embraced nor rejected

Breakout 3 – David Boyd

Volume 6, Reference Architecture

- Need to put the conceptual view in context of stakeholders and their concerns
  - Reflect domain needs and requirements
  - Consider new diagram
- Need to address Data Management
  - Own fabric?
  - Subrole of management (Data and System subroles)
- System Orchestrator issues
  - Business or Technical
  - Relates back to compromises and multiple thoughts in V1
  - Add Finances/Costs and also Service Level requirements
  - Consider renaming
- Find place to represent Provenance and Pedigree activities
- Add 4th to Security and Privacy (Accounting? Find definitions)
- Need text on non-linear flow – Consider mapping use case flows through architecture.

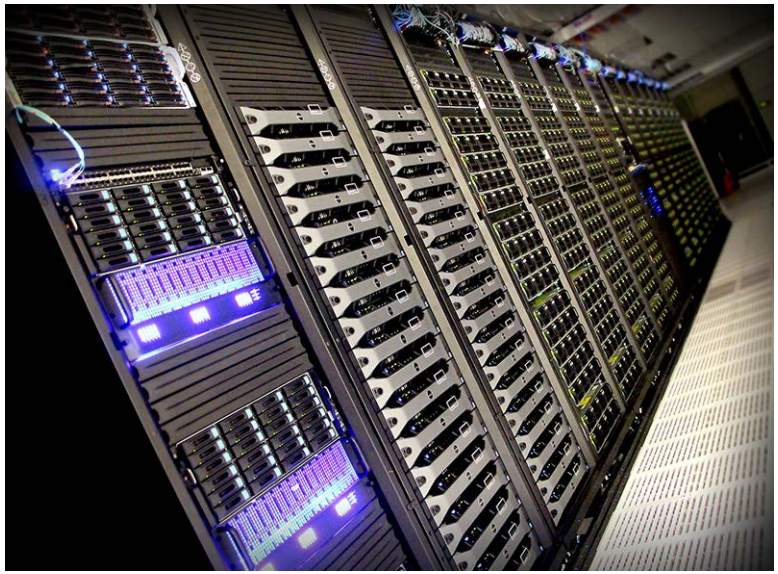
Breakout 3 – Gregor von Laszewski (laszewski@gmail.com)

Discussions Vol.8

- Volume 8 reference implementation implies that a service layer exist as part of the system orchestrator
- REST is one architectural approach to be used in a reference implementation, other could exist.
  - Reminder REST is one technology to implement SaaS
  - Add Section to stress that choice.
  - Move REST related terminology possibly in the appendix
- JSON format proved as very advantageous.
  - Some asked should we use more complex specifications such as xml
- Object:
  - Possibly using “Resource” instead of “Object” but that binds us to REST
  - Use Model? This may provide its own challenges
- Remember Vol 8 must be supporting reference implementation

Cloudmesh Tutorial

- Cloudmesh Tutorial at PEARC17
- **Monday**, July 10 • 9:00am - 12:30pm, New Orleans
- Advanced Tutorial to learn how to create virtual clusters with the help of Cloudmesh on NSF sponsored Comet
- Comet
  - > 15000 Cores
  - > 1500 Servers
  - 7 PB storage



Breakout 4 – Russell Reinsch

Volume 7, Standards Roadmap

- The need to include more pointers in the docs to other relevant sources
- Our roadmap needs to have a “predictive aspect”, not just a catalog
- Perhaps add a column structure for quick search and relevance for the reader
- Need to articulate the spectrum of standards (defacto standard, normative or informative by ISO definition, etc.)

Breakout 4 – Russell Reinsch

Volume 9, Adoption and Modernization

- Need to address governance in more detail
- Address cloud integration