Overview NIST Big Data Working Group Activities and Big Data Architecture Framework (BDAF) by UvA

Yuri Demchenko
SNE Group, University of Amsterdam

Big Data Analytics Interest Group
17 September 2013, 2nd RDA Plenary
Outline

• Overview NIST Big Data Working Group (NBD-WG) activities and deliverables
• Proposed Big Data Architecture Framework (BDAF)
  – Data Models and Big Data Lifecycle
  – Big Data Infrastructure (BDI)
• Discussion: Liaison and information exchange with NIST BD-WG

Disclaimer: Presented here information about NIST Big Data Working Group (NBD-WG) and images from the NBD-WG working documents are not official position of the NBD-WG and are solely the authors opinion.
NIST Big Data Working Group (NBD-WG)

• Deliverables target – September 2013
  – 26 September – initial draft documents
  – 30 September – Workshop and F2F meeting

• Activities: Conference calls every day 17-19:00 (CET) by subgroup - [http://bigdatawg.nist.gov/home.php](http://bigdatawg.nist.gov/home.php)
  – Big Data Definition and Taxonomies
  – Requirements (chair: Geoffrey Fox, Indiana Univ)
  – Big Data Security
  – Reference Architecture
  – Technology Roadmap

• BigdataWG mailing list and useful documents
  – Big Data Reference Architecture [http://bigdatawg.nist.gov/_uploadfiles/M0226_v2_1885676266.docx](http://bigdatawg.nist.gov/_uploadfiles/M0226_v2_1885676266.docx)
  – Requirements for 21 usecases [http://bigdatawg.nist.gov/_uploadfiles/M0224_v1_1076079077.xlsx](http://bigdatawg.nist.gov/_uploadfiles/M0224_v1_1076079077.xlsx)
NIST Proposed Reference Architecture (before July 2013)

- Obviously not data centric
- Doesn’t make data (lifecycle) management clear
[ref] NIST Big Data WG mailing list discussion
http://bigdatawg.nist.gov/_uploadfiles/M0010_v1_6762570643.pdf
Big Data Ecosystem Reference Architecture (By Microsoft) [ref] – Initial contribution July 2013

[ref] Big Data Ecosystem Reference Architecture (Microsoft)
http://bigdatawg.nist.gov/_uploadfiles/M0015_v1_1596737703.docx
NIST Reference Architecture version 0.0 (August 2013)
Big Data Architecture Framework (BDAF) by the University of Amsterdam

- Big Data definition: from 5+1Vs to 5 parts
- Big Data Architecture Framework (BDAF) components
Improved: 5+1 V’s of Big Data

**Volume**
- Terabytes
- Records/Arch
- Tables, Files
- Distributed

**Variety**
- Structured
- Unstructured
- Multi-factor
- Probabilistic
- Linked
- Dynamic
- Changing data
- Changing model
- Linkage

**Velocity**
- Batch
- Real/near-time
- Processes
- Streams
- Correlations
- Statistical
- Events
- Hypothetical

**Veracity**
- Trustworthiness
- Authenticity
- Origin, Reputation
- Availability
- Accountability

**Value**

**Generic Big Data Properties**
- Volume
- Variety
- Velocity

**Acquired Properties (after entering system)**
- Value
- Veracity
- Variability
Big Data Definition: From 5+1V to 5 Parts (1)

(1) Big Data Properties: 5V
   - Volume, Variety, Velocity, Value, Veracity
   - Additionally: Data Dynamicity (Variability)

(2) New Data Models
   - Data Lifecycle and Variability
   - Data linking, provenance and referral integrity

(3) New Analytics
   - Real-time/streaming analytics, interactive and machine learning analytics

(4) New Infrastructure and Tools
   - High performance Computing, Storage, Network
   - Heterogeneous multi-provider services integration
   - New Data Centric (multi-stakeholder) service models
   - New Data Centric security models for trusted infrastructure and data processing and storage

(5) Source and Target
   - High velocity/speed data capture from variety of sensors and data sources
   - Data delivery to different visualisation and actionable systems and consumers
   - Full digitised input and output, (ubiquitous) sensor networks, full digital control
Refining Gartner definition

- Big Data (Data Intensive) Technologies are targeting to process (1) high-volume, high-velocity, high-variety data (sets/assets) to extract intended data value and ensure high-veracity of original data and obtained information that demand cost-effective, innovative forms of data and information processing (analytics) for enhanced insight, decision making, and processes control; all of those demand (should be supported by) new data models (supporting all data states and stages during the whole data lifecycle) and new infrastructure services and tools that allows also obtaining (and processing data) from a variety of sources (including sensor networks) and delivering data in a variety of forms to different data and information consumers and devices.

(1) Big Data Properties: 5V
(2) New Data Models
(3) New Analytics
(4) New Infrastructure and Tools
(5) Source and Target
Big Data Nature: Origin and consumers (target)

### Big Data Origin
- Science
- Telecom
- Industry
- Business
- Living Environment, Cities
- Social media and networks
- Healthcare

### Big Data Target Use
- Scientific discovery
- New technologies
- Manufacturing, processes, transport
- Personal services, campaigns
- Living environment support
- Healthcare support
# Big Data Nature: Origin and Consumers (Targets)

<table>
<thead>
<tr>
<th>Science Discovery</th>
<th>New Technology</th>
<th>Manufactur Transport</th>
<th>Personal services, campaigns</th>
<th>Living Environment, Infrastructure, Utility</th>
<th>Healthcare support</th>
</tr>
</thead>
<tbody>
<tr>
<td>Science</td>
<td>++++</td>
<td>+++</td>
<td>+</td>
<td>-</td>
<td>++</td>
</tr>
<tr>
<td>Telecom</td>
<td>+</td>
<td>+++</td>
<td>++</td>
<td>+</td>
<td>+++</td>
</tr>
<tr>
<td>Industry</td>
<td>++</td>
<td>+++</td>
<td>++++</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Business</td>
<td>+</td>
<td>+++</td>
<td>++</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>Living environment, Cities</td>
<td>++</td>
<td>++</td>
<td>++</td>
<td>++</td>
<td>++++</td>
</tr>
<tr>
<td>Social media, networks</td>
<td>+</td>
<td>++</td>
<td>-</td>
<td>+++</td>
<td>++</td>
</tr>
<tr>
<td>Healthcare</td>
<td>+++</td>
<td>++</td>
<td>-</td>
<td>-</td>
<td>++</td>
</tr>
</tbody>
</table>

Moving to Data-Centric Models and Technologies

• Current IT and communication technologies are host based or host centric
  – Any communication or processing are bound to host/computer that runs software
  – Especially in security: all security models are host/client based

• Big Data requires new data-centric models
  – Data location, search, access
  – Data variability and lifecycle
  – Data integrity and identification
  – Data centric security and access control
Defining Big Data Architecture Framework

- Existing attempts don’t converge to consistent view: ODCA, TMF, NIST
  - See http://bigdatawg.nist.gov/_uploadfiles/M0055_v1_7606723276.pdf

- Big Data Architecture Framework (BDAF) by UvA
  Architecture Framework and Components for the Big Data Ecosystem. Draft Version 0.2

- Architecture vs Ecosystem
  - Big Data undergo a number of transformations during their lifecycle
  - Big Data fuel the whole transformation chain
    - Data sources and data consumers, target data usage
  - Multi-dimensional relations between
    - Data models and data driven processes
    - Infrastructure components and data centric services

- Architecture vs Architecture Framework (Stack)
  - Separates concerns and factors
    - Control and Management functions, orthogonal factors
  - Architecture Framework components are inter-related
Big Data Architecture Framework (BDAF) for Big Data Ecosystem (BDE)

(1) Data Models, Structures, Types
   - Data formats, non/relational, file systems, etc.

(2) Big Data Management
   - Big Data Lifecycle (Management) Model
     • Big Data transformation/staging
   - Provenance, Curation, Archiving

(3) Big Data Analytics and Tools
   - Big Data Applications
     • Target use, presentation, visualisation

(4) Big Data Infrastructure (BDI)
   - Storage, Compute, (High Performance Computing,) Network
   - Sensor network, target/actionable devices
   - Big Data Operational support

(5) Big Data Security
   - Data security in-rest, in-move, trusted processing environments
Big Data Architecture Framework (BDAF) – Aggregated – Relations between components (2)

<table>
<thead>
<tr>
<th>Col: Used By</th>
<th>Data Models &amp; Structures</th>
<th>Data Managmnt &amp; Lifecycle</th>
<th>BigData Infrastr &amp; Operations</th>
<th>BigData Analytics &amp; Applicatn</th>
<th>BigData Security</th>
</tr>
</thead>
<tbody>
<tr>
<td>Row: Requires This</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data Models &amp; Structures</td>
<td></td>
<td>+</td>
<td>++</td>
<td>+</td>
<td>++</td>
</tr>
<tr>
<td>Data Managmnt &amp; Lifecycle</td>
<td>++</td>
<td></td>
<td>++</td>
<td>++</td>
<td>++</td>
</tr>
<tr>
<td>BigData Infrastruct &amp; Operations</td>
<td>+++</td>
<td>+++</td>
<td></td>
<td>++</td>
<td>+++</td>
</tr>
<tr>
<td>BigData Analytics &amp; Applications</td>
<td>++</td>
<td>+</td>
<td>++</td>
<td></td>
<td>++</td>
</tr>
<tr>
<td>BigData Security</td>
<td>+++</td>
<td>+++</td>
<td>+++</td>
<td>+</td>
<td></td>
</tr>
</tbody>
</table>
Big Data Ecosystem: Data, Lifecycle, Infrastructure

Data Source

Data Collection & Registration → Data Filter/Enrich, Classification → Data Analytics, Modeling, Prediction → Data Delivery, Visualisation → Consumer

Big Data Target/Customer: Actionable/Usable Data
Target users, processes, objects, behavior, etc.

Big Data Source/Origin (sensor, experiment, logdata, behavioral data)

Big Data Analytic/Tools

Storage General Purpose → Compute General Purpose → High Performance Computer Clusters → Storage Specialised Databases Archives (analytics DB, in memory, operational)

Federated Access and Delivery Infrastructure (FADI)

Intercloud multi-provider heterogeneous Infrastructure

Security Infrastructure
Network Infrastructure Internal
Infrastructure Management/Monitoring

Data categories: metadata, (un)structured, (non)identifiable
Data Transformation/Lifecycle Model

- Does Data Model changes along lifecycle or data evolution?
- Identifying and linking data
  - Persistent identifier
  - Traceability vs Opacity
  - Referral integrity

Common Data Model?
- Data Variety and Variability
- Semantic Interoperability
Scientific Data Lifecycle Management (SDLM) Model

Data Lifecycle Model in e-Science

Data discovery

Data collection and filtering

Data analysis

Data sharing/Data publishing

End of project

Data re-purpose

Data Curation (including retirement and clean up)

Data archiving

Raw Data
Experimental Data

Structured Scientific Data

Data linkage to papers

Data archiving

Data linkage

to papers

Data re-cycling

Open Public Use

Data Links

Metadata & Mngnt

Data linkage Issues

- Persistent Identifiers (PID)
- ORCID (Open Researcher and Contributor ID)
- Lined Data

Data Clean up and Retirement

- Ownership and authority
- Detainment

User

Researcher

Project/Experiment Planning

Data Re-purpose

Data Re-purpose

Data Re-purpose

Data Re-purpose

Data Re-purpose

Data Re-purpose

Data Re-purpose

Data Re-purpose

Data Re-purpose

Data Re-purpose

Data Re-purpose

Data Re-purpose

Data Re-purpose

Data Re-purpose

Data Re-purpose

Data Re-purpose

Data Re-purpose

Data Re-purpose

Data Re-purpose
Evolutional/Hierarchical Data Model

- Common Data Model?
- Data interlinking?
- Fits to Graph data type?
- Metadata

- Referrals
- Control information
- Policy
- Data patterns
Additional Information

- Existing proposed Big Data architectures
Industry Initiatives to define Big Data (Architecture)

- Open Data Center Alliance (ODCA) Information as a Service (INFOaaS)
- TMF Big Data Analytics Reference Architecture
- Research Data Alliance (RDA)
  - All data related aspects, but not Infrastructure and tools
- LexisNexis HPCC Systems
ODCA INFOaaS – Information as a Service

- Using integrated/unified storage
  - New DB/storage technologies allow storing data during all lifecycle

[ref] Open Data Center Alliance Master Usage model: Information as a Service, Rev 1.0.
http://www.opendatacenteralliance.org/docs/Information_as_a_Service_Master_Usage_Model_Rev1.0.pdf
ODCA Example INFOaaS Architecture

- Core Data and Information Components
- Data Integration and Distribution Components

• Presentation and Information Delivery Components
• Control and Support Components
TMF Big Data Analytics Architecture

Real-Time Stream Processing
Complex Event Processor, Alerts & Triggers

Data Analysis
Metrics Calculation, Data Modeling

Data Management
Transformation, Correlation, Enrichment, Manipulation

Data Ingestion
Integration, Import, Data Formatting

Data Sources
Network, OSS, BSS, Social Networks, ...

LexisNexis Vision for Data Analytics Supercomputer (DAS) [ref]

LexisNexis HPCC System Architecture

ECL – Enterprise Data Control Language
THOR Processing Cluster (Data Refinery)
Roxie Rapid Data Delivery Engine

IBM GBS Business Analytics and Optimisation (2011).
https://www.ibm.com/developerworks/mydeveloperworks/files/basic/anonymous/api/library/48d92427-47d3-4e75-b54c-b6acfb608c0/document/aa78f77c-0d57-4f41-a923-50e5c6374b6d/media&ei=yrknUbMNM_liwKQhoCQBQ&usg=AFQjCNF_Xu6aifcAhlF4266xXNhKfKaTLw&sig2=j8JiFV_md5DnzfQl0spVrg&bvm=bv.42768644,d.CGE